Novel antibiotics from a ‘white box’ 2D structural fingerprint decision tree

Gareth Williams

Wolfson Centre for Age-Related Diseases, King’s College London, UK

Email: Gareth.2.williams@kcl.ac.uk

Abstract

Drug repurposing is the application of existing therapeutics for indications for which they were not developed. This approach has opened up drug discovery to the academic lab, government agencies and not-for-profit organisations. Repurposing can tap into the vast amount of publicly available data on drug activities in the form of gene expression profiling, cell culture assay profiles and gene interaction patterns. One recent endeavour sought to discover novel antibiotics through a ‘black box’ neural network algorithm based on compound structure bond connectivity. Training and validation experiments on compound datasets with known antimicrobial activity facilitated the discovery of the antibiotic properties of the diabetic medication halicin. In this work we show that a simple ‘white box’ decision tree based on 2D PubChem structural fingerprints correlates with the neural network approach and also picks out halicin as an antibiotic candidate with an explicit pattern for the assignment.

Introduction

The harnessing of the wider therapeutic potential of the existing therapeutics through repurposing has the potential to accelerate the drug discovery pipeline and widen participation beyond the pharmaceutical industry. Access to the safety profiles and extensive prescription data may allow for preclinical phase I and II testing, the stumbling blocks for most novel entities, to be bypassed. Approaches to repurposing range from direct epidemiological correlation analyses, such as the observation of low Parkinson’s disease incidence correlating with salbutamol use (Mittal, Bjornevik et al. 2017). At the other extreme of mechanistic detail, repurposing can proceed through the emergence of specific targets for which there are specific drugs as in a recent study into SARS-CoV-2 based on compiling evidence for interactions between viral and host cell proteins (Gordon, Jang et al. 2020). The host proteins were then queried against drug interaction databases, such as the EBI ChEMBL resource (www.ebi.ac.uk/chembl). An intermediate methodology taps into large scale quantitative drug associated activity data such as gene expression perturbation. This approach has revealed the intriguing transcriptional similarity between disparate antiviral compounds (Killick, Ballard et al. 2020). Transcription-based drug repurposing gaining interest in the field of neurodegenerative diseases (Ballard, Aarsland et al. 2020). Another multiparameter descriptor of compound activity is provided by the NCI Compare (Zaharevitz, Holbeck et al. 2002), which has been applied to the current COVID19 pandemic and shown to link disparate, chemically unrelated, drugs with reported antiviral activities (Naasani 2021).

Compound activities are ultimately of course encoded in their structures and this has led to an interest in the quantification of structural similarity and how this relates to similarities in activity. Unlike DNA and protein comparison, which can be scored through sequence alignment, compound similarity scoring has been done through the compilation of the relative spatial arrangement of chemical moieties or pharmacophores. Rapid compound comparison incorporating structural flexibility is offered through the ROCS algorithm provided by OpenEye (www.eyesopen.com). Alternatively, linear fingerprints based on categorical calls on the presence of a catalogue of chemical building blocks facilitate a simple scoring methodology that does not suffer from the conformational variability of pharmacophores. PubChem (pubchem.ncbi.nlm.nih.gov) hosts fingerprints for their collection of over 100 million compounds and these comprise information on 881 distinct molecular fragments (ftp://ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt). The ChemAxon (docs.chemaxon.com) provides freeware to generate a variety of fingerprints usually spanning 1024 fragments. Fingerprinting is usually used for speedy compound database querying either for overall similarity or substructure searches.

Recently an artificial intelligence (AI) study repurposed halicin as an antibiotic (Stokes, Yang et al. 2020). In this study US Food and Drug Administration approved drugs and natural products were screened for the inhibition of *E. coli* growth and 120 out of the 2,335 exhibited inhibitory activity. Directed-message passing deep neural network (DMPNN) modelling was applied to a training set of compounds with known antibacterial activities. The method relies on molecular features that are learned during training and supplemented by RDKit molecular features and effectively segregates compounds based on antibacterial activity. A collection of 6,111 preclinical compounds comprising the Drug Repurposing Hub (DRH) was then filtered to remove compounds in the training set to leave 4,496 and ranked for predicted antibacterial activity with the DMPNN score. Of the top 99 ranked compounds 51 were shown to display antibacterial in the *E. coli* growth inhibition assay. Halicin, an anti-diabetic drug, was prioritised based on it having a low structural similarity to the original training set compound and its low toxicity profile. Subsequently halicin was shown to be a broad-spectrum antibiotic having activity against antibiotic tolerant cells. The rank position of halicin (89th out of 4,496) was shown to be higher than that achieved by other neural network and random forest fingerprint-based approaches (highest rank method placing halicin at 273 or the top 6%).These approaches come under the heading of 'black box' prediction i.e. the structural features driving activity are obscure at best. It is interesting therefore to see to what extent similar results emerge with a 'white box' methodology. The simplest such is a decision tree (DT) based on 2D structural fingerprints. It is shown below that a simple DT approach trained on the same data returns a good correlation with the AI ranking of the DRH database and further halicin is in the top 6% as opposed to the top 2% for the AI protocol and 6% for the alternative scoring systems reported in the study.

Methods and Results

To implement a DT the PubChem fingerprints for the compounds under consideration were downloaded from the NCBI ftp site ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/SDF/ after mapping compound names to their PubChem compound ID (CID) through interrogation of synonym tables ftp.ncbi.nlm.nih.gov/pubchem/Compound/Extras/CID-Synonym-filtered.gz. The fingerprints, as mentioned above, consist of 881 element binary strings encoding the presence of structure elements defined in terms of the 2D compound connectivity and element composition. Of the 2,335 compounds in the training dataset 2,238 were unambiguously identified and this split into 118 active and 2119 inactive entries. In terms of the PubChem fingerprinting there is equivalence between tobramycin and dibekacin, which are assigned active and inactive status respectively. We therefore dropped dibekacin from the list.

An optimal DT constructed based on the activity assignment of the training dataset with branches corresponding to yes/no calls on the presence of a given fingerprint feature. Branching is optimised to maximise the entropy loss. Here, high entropy corresponds to a mix of active and inactive compounds and zero entropy corresponds to only actives or in-actives present in each branch. For example, if a dataset of *D* members is split into two sets *N* and *M* with yes/no activity counts of (*n*, *N*-*n*) and (*m*, *M*-*m*) then the weighted entropy sum is:

$$\frac{N}{D}\left(-\frac{n}{N}log\left(\frac{n}{N}\right)-\left(1-\frac{n}{N}\right)log\left(1-\frac{n}{N}\right)\right)+\frac{M}{D}\left(-\frac{m}{M}log\left(\frac{m}{M}\right)-\left(1-\frac{m}{M}\right)log\left(1-\frac{m}{M}\right)\right)$$

If the split separates yes/no calls exactly then the resulting entropy is zero of course. Zero entropy nodes are termini of the tree.

Applying this methodology to the antibiotic training set the resulting tree is constituted by 92 pure terminal nodes (42 antibacterial and 50 non-antibacterial) with 19 layers and by construction segregates the 118 active compounds from the 2119 inactive compounds. The antibacterial node distribution is skewed with 24 nodes occupied by one compound instance and a Shannon entropy of 3.16.

It was then of interest to see how this DT segregates the DRH compounds. Here, a direct comparison can be made with the ranks reported for the methodologies reported in the AI paper. Applying the tree scoring to the DRH set we see a clear correlation with the DMPNN score, see Figure 1.

Taking the reported compound rank as the discriminator and a positive DT node score as a successful result the comparison between the approaches can be cast as a ROC area under the curve analysis (ROC-AUC). For the results are shown in Figure 2. Interestingly, the DT has the closest correlation with DMPNN scoring.

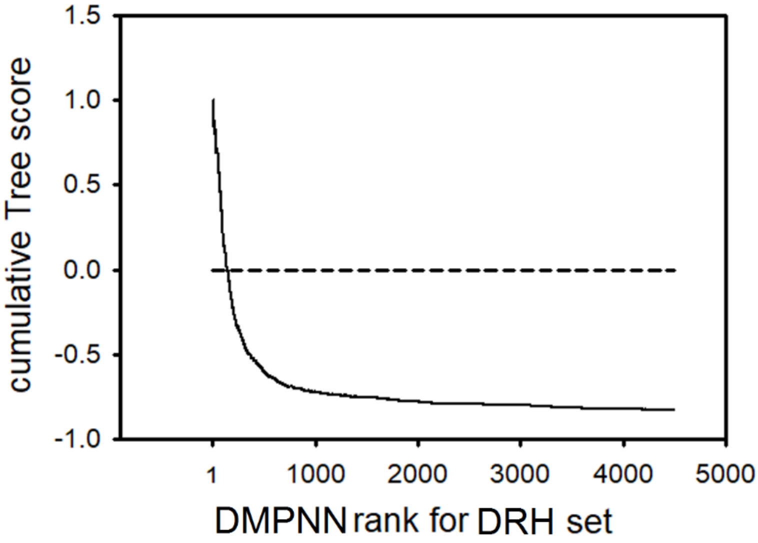


Figure 1. The cumulative tree score for the DRH compounds ranked according to the DMPNN score. There is a clear enrichment of positive tree scores for highly ranked compounds, thus showing largely equivalent outcomes for the black and 'white box' approaches.

ROC-AUC	Methodology
0.73	DMPNN
0.70	Without RDKit features
0.72	Just RDKit features
0.68	Feed forward DMP with Morgan fingerprints
0.54	Random Forest with Morgan fingerprints
0.66	Support Vector Machine with Morgan fingerprints

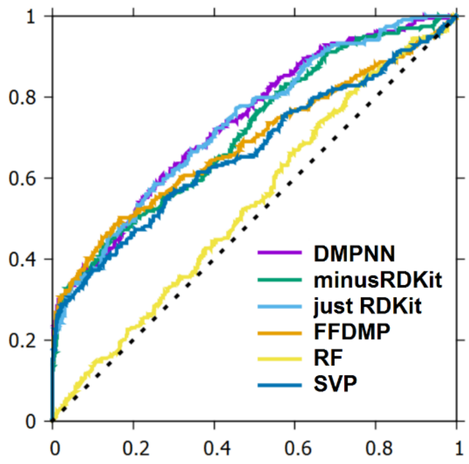
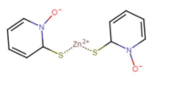


Figure 2. The comparison of the DT terminal node scores with the various methodology rankings for the DRH compounds that were not part of the training stage. The DMPNN rankings appears to most closely correlate with the DT. The other methodologies score slightly worse apart from that of the random forest approach using Morgan fingerprints.

In total 239 or 5.7% of the DRH compounds return a DT score of one and are predicted to be antibiotics. In the top 25 DMPNN ranked compounds only two have negative DT scores. And in the top 50 84% are assigned positive scores. Halicin is also assigned antibacterial status and this compares to a ranking in the top 2% for DMPNN and the top 6% for the next highest-ranking scoring scheme (RDKit features only) not incorporating learned features. Of the top 99 DMPNN ranked compounds 50 displayed antibacterial activity in assays, but the compounds were not given explicitly, consequently the DT score for the active compounds can't be determined here.

It is of interest to examine the DT route to halicin being assigned antibacterial activity. This is shown in Figure 3 together with the 2D structure. There is only one entry in the training set with this pattern and this might explain the poor ranking of halicin in the random forest approach where trees are based on subsets of compounds, see Figure 2.

O=C-C-C-N	no
OC1C(N)CCCC1	no
O-N-C-C	yes
O=C-C-C-C-C-C	no
N(~C)(~C)(~C)	no
O(~C)(~C)	no
O(~C)(~H)	no
>=2S	yes

CID	Name	Structure	Chembl
26041	Zinc Pyrithione		CHEMBL3392049

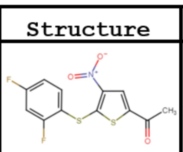
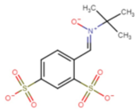
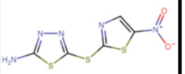
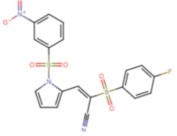
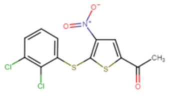
CID	Name	Structure	Chembl
2819993	882257-11-6		CHEMBL2159495
6440181	Cerovive		CHEMBL1627056
11837140	Halicin		CHEMBL510038
44607965	AMZ30		CHEMBL1550905
46931953	1247819-59-5		CHEMBL2159498

Figure 3. Compounds sharing halicin's antibacterial fingerprint pattern. The DT pattern assigning halicin to antibacterial status is shown at the left. The fingerprint pattern occurs once in the training database and this corresponds to Zinc Pytrithione, shown top at the right. In the DRH database there are four compounds in addition to halicin harbouring the same fingerprint pattern, shown bottom at right.

A relatively comprehensive list of antimicrobial agents can be obtained from the EBI hosted chemical properties database ChEMBL (www.ebi.ac.uk/chembl/). The IC50 activities database (www.ebi.ac.uk/chembl/g/#browse/activities) when filtered for antimicrobial activity ChEMBL assay assignment and with reported activity values results in a list of 2237 compounds, 78 of which are in the DT training database. Out of the 2159 a total of 140 or 6.5% are assigned antibacterial activity based on the tree scoring. This is a slightly higher hit rate than for the DRH, which is not enriched for antibacterials and therefore the increased hit rate for the ChEMBL set is not surprising. However, the marginal increase in antibacterial assignment relative to a non-enriched compound set does not constitute a predictive methodology for antibacterial activity. It would be of interest to assess the performance of the DMPNN technique in this context.

Discussion

Repurposing therapeutics is an increasingly attractive drug discovery route and can proceed via an examination of drug activity relative to disease state. For example, a new intervention target may emerge for which there is an 'off-the-shelf' therapeutic option or some high content disease state descriptor, such as provided by transcription profiling, can be linked to a similar drug action descriptor. Also, alternative drug candidates can be sourced based on their sharing properties with existing therapeutics. This can be in the form of activity profiles or based on structure. Structural fingerprints provide a linear representation of structure that can be used to quantify structural similarity. Fingerprints can also be used to construct DTs when there are sufficient examples of diverse compounds sharing activity. Stochastically determined populations of DTs with small layer counts can be combined into random forests to deal with fragility issues in single DT based approaches. Recently, a sophisticated neural network approach based on compound bond connectivity has been developed and deployed in the search for novel antibacterial agents. This DMPNN methodology was shown to be effective with 50% of the top 100 candidates subsequently shown to be antibacterials. A notable candidate was halicin, a diabetic medication, which showed relatively low toxicity and was therefore the top repurposing candidate in the study. In the present work, a simple DT approach is shown to recapitulate some to the results of the DMPNN. In contrast to random forest approaches and neural network schemes, the DT framework provides a 'white box' description of the key features imparting antibacterial activity.

References

Ballard, C., D. Aarsland, J. Cummings, J. O'Brien, R. Mills, J. L. Molinuevo, T. Fladby, G. Williams, P. Doherty, A. Corbett and J. Sultana (2020). "Drug repositioning and repurposing for Alzheimer disease." Nat Rev Neurol **16**(12): 661-673.

Gordon, D. E., G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney, T. A. Tummino, R. Huettenhain, R. M. Kaake, A. L. Richards, B. Tutuncuoglu, H. Foussard, J. Batra, K. Haas, M. Modak, M. Kim, P. Haas, B. J. Polacco, H. Braberg, J. M. Fabius, M. Eckhardt, M. Soucheray, M. J. Bennett, M. Cakir, M. J. McGregor, Q. Li, B. Meyer, F. Roesch, T. Vallet, A. Mac Kain, L. Miorin, E. Moreno, Z. Z. C. Naing, Y. Zhou, S. Peng, Y. Shi, Z. Zhang, W. Shen, I. T. Kirby, J. E. Melnyk, J. S. Chorba, K. Lou, S. A. Dai, I. Barrio-Hernandez, D. Memon, C. Hernandez-Armenta, J. Lyu, C. J. P. Mathy, T. Perica, K. B. Pilla, S. J. Ganesan, D. J. Saltzberg, R. Rakesh, X. Liu, S. B. Rosenthal, L. Calviello, S. Venkataramanan, J. Liboy-Lugo, Y. Lin, X. P. Huang, Y. Liu, S. A. Wankowicz, M. Bohn, M. Safari, F. S. Ugur, C. Koh, N. S. Savar, Q. D. Tran, D. Shengjuler, S. J. Fletcher, M. C. O'Neal, Y. Cai, J. C. J. Chang, D. J. Broadhurst, S. Klippsten, P. P. Sharp, N. A. Wenzell, D. Kuzuoglu, H. Y. Wang, R. Trenker, J. M. Young, D. A. Cavero, J. Hiatt, T. L. Roth, U. Rathore, A. Subramanian, J. Noack, M. Hubert, R. M. Stroud, A. D. Frankel, O. S. Rosenberg, K. A. Verba, D. A. Agard, M. Ott, M. Emerman, N. Jura, M. von Zastrow, E. Verdin, A. Ashworth, O. Schwartz, C. d'Enfert, S. Mukherjee, M. Jacobson, H. S. Malik, D. G. Fujimori, T. Ideker, C. S. Craik, S. N. Floor, J. S. Fraser, J. D. Gross, A. Sali, B. L. Roth, D. Ruggero, J. Taunton, T. Kortemme, P. Beltrao, M. Vignuzzi, A. Garcia-Sastre, K. M. Shokat, B. K. Shoichet and N. J. Krogan (2020). "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing." Nature.

Killick, R., C. Ballard, P. Doherty and G. Williams (2020). "Transcription-based drug repurposing for COVID-19." Virus Res **290**: 198176.

Mittal, S., K. Bjornevik, D. S. Im, A. Flierl, X. Dong, J. J. Locascio, K. M. Abo, E. Long, M. Jin, B. Xu, Y. K. Xiang, J. C. Rochet, A. Engeland, P. Rizzu, P. Heutink, T. Bartels, D. J. Selkoe, B. J. Caldarone, M. A. Glicksman, V. Khurana, B. Schule, D. S. Park, T. Riise and C. R. Scherzer (2017). "beta2-Adrenoreceptor is a regulator of the alpha-synuclein gene driving risk of Parkinson's disease." Science **357**(6354): 891-898.

Naasani, I. (2021). "COMPARE Analysis, a Bioinformatic Approach to Accelerate Drug Repurposing against Covid-19 and Other Emerging Epidemics." SLAS Discov **26**(3): 345-351.

Stokes, J. M., K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay and J. J. Collins (2020). "A Deep Learning Approach to Antibiotic Discovery." Cell **180**(4): 688-702 e613.

Zaharevitz, D. W., S. L. Holbeck, C. Bowerman and P. A. Svetlik (2002). "COMPARE: a web accessible tool for investigating mechanisms of cell growth inhibition." J Mol Graph Model **20**(4): 297-303.