Managing Antibiotics Research Data with CDD Vault, a Drug Discovery Collaborative Platform

Mariana Vaschetto*
Susana Tomasio
Kellan Gregory

* mariana@collaborativedrug.com

ABSTRACT

The COVID-19 pandemia has brought into focus the need to be prepared for emerging threats. Arguably one of the most significant concerns in the area is antibiotics resistance. There is a vast amount of data in the Antibiotic field that has been collected over the years and can be used as the springboard to new Drug Discovery projects. Literature data is noisy and need to be curated and trusted to be useful.

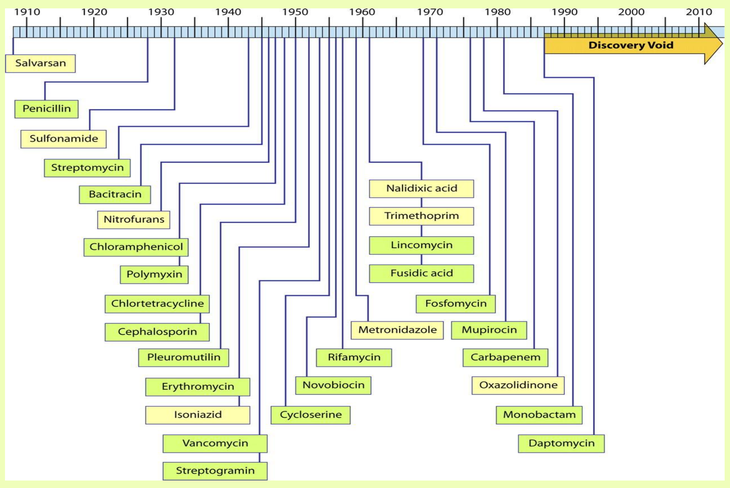
In addition, successful projects such as the COVID Moonshot with a worldwide approach to crowd sourcing science ideas has shown that collaborative efforts are key to accelerate the delivery of new therapies[1].

We present an application of the CDD Vault research data management platform combining the benefits of using curated antibacterial (Gram-negative) research data in combination with machine learning methods in a collaborative platform to tackle resistance.

We use CDD Vault®, a hosted platform designed to archive, mine, and securely share research data across multiple sites and verticals, to accelerate the discovery process by sharing knowledge and build models together based on the past and collective know-how.

Introduction

During the last 30 years, there has been a lack of new antibiotics coming to the market, as it is shown n this figure first published by Lynn Silver in 2011,



Lynn L. Silver Clin. Microbiol. Rev. 2011; 24: 71-109

There is a latent risk of a serious outbreak of antibiotics resistant bacteria due to,

- Lack of new class of antibiotics
- Increased resistance to known antibiotics

Collaboration in antibiotics research could be key to replenish the antibiotics pipeline. The collaborative nature of the CDD Vault informatics platform has proven to be ideal for this type of large, decentralized collaborations that gather information throughout the whole drug discovery process. CDD Vault is capable of storing and organizing chemical and biological data coming from different sources. Having a centralized, single source of truth enables distributed teams to analyse and share information in real time, accelerating progress and increasing the chance of success.
See for instance, Dr. Lizbe Koekemoer of Oxford University describing COVID Moonshot project here:

https://www.collaborativedrug.com/covid-moonshot-cdd-supports-worldwide-effort-find-therapy/

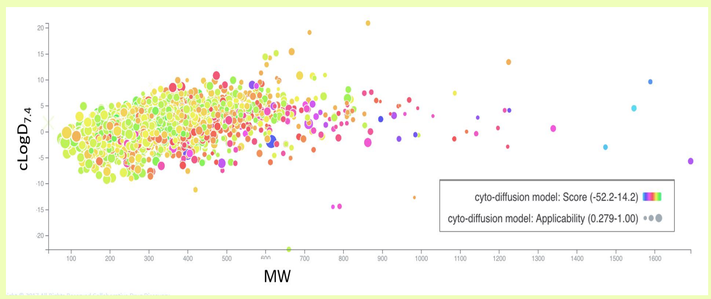
DRUG REPURPOSING: ANTIBIOTICS

Building a Model

Training set
- 53 compounds of known drugs
- GN cytoplasmic diffused antibacterial compounds
- cLogD between -3 and 4
- MW < 600

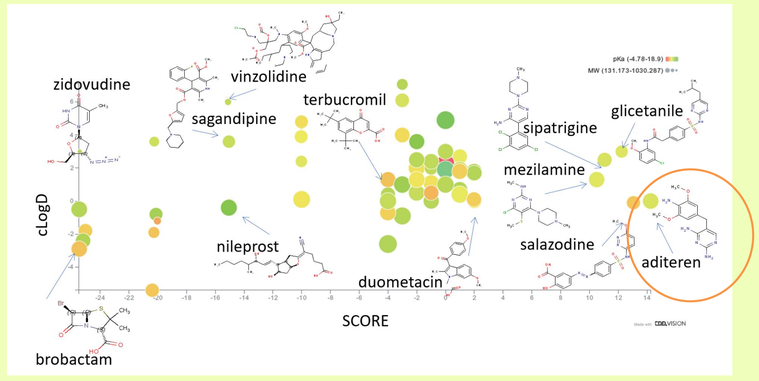
The cyto-diffusion model was built in CDD Vault as described using a Bayesian algorithm and FCFP6 fingerprints.

The model was subsequently applied to 4604 non-antibiotics compounds from the CMC database.



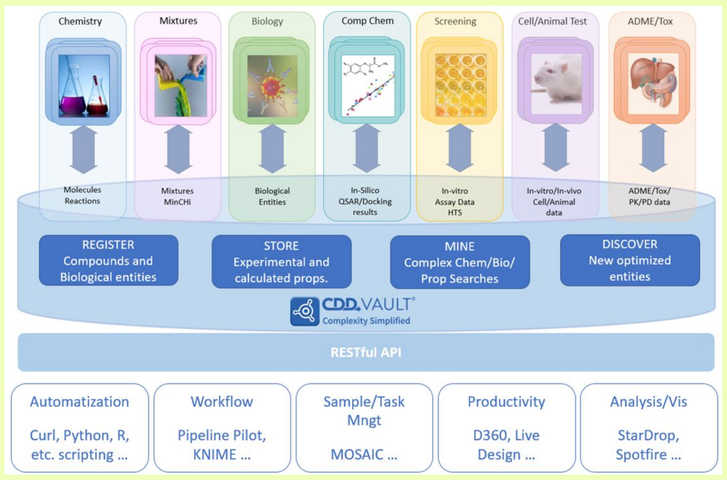
Despite the heterogeneity of the data, the parameters show a good model for scoring cyto-diffused compounds,
Score: -52.2 - 14.2
Applicability: 0.279-1.00

A selection of the best scored compounds with potential for repurposing shows:



Aditeren is a pyrimidinediamine derivative and dihydrofolate reductase inhibitor related to trimethoprim, with bactericide and diuretic activities. Aditeren has never been marketed.

CDD Vault



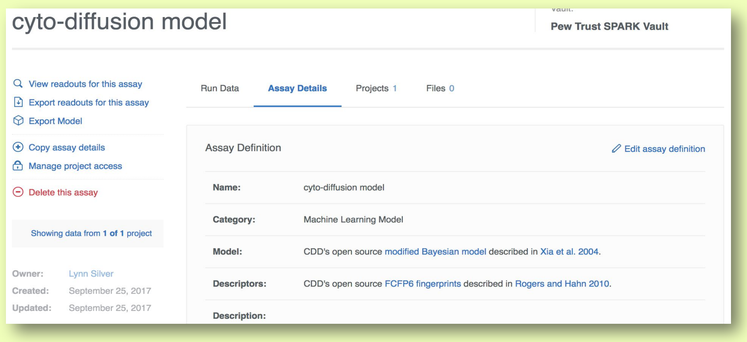
References

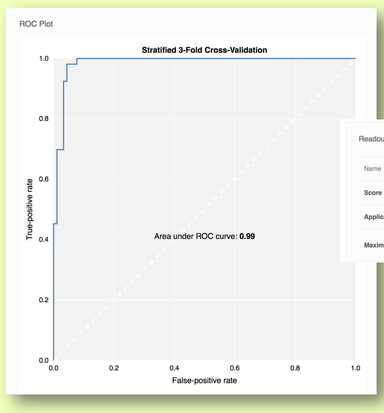
1. https://www.biorxiv.org/content/10.1101/2020.10.29.339317v1
2. Challenges of Antibacterial Discovery (Silver, L. Clin Microbiol Rev. 2011)
3. https://www.collaborativedrug.com/public-access/
4. Roger, D. and Hahn, M., *J. Chem. Inf. Model.* 2010, 50, 5, 742–754
5. Xia, X. et al. *J. Med. Chem.* 2004, 47, 18, 4463−4470

Methods

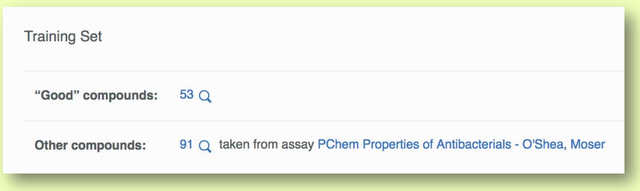
We use the machine learning model extension in CDD Vault.

CDD Vault uses FCFP6 structural fingerprints to build a Bayesian statistical model. FCFP6 fingerprints are described in Rogers and Hahn[4].
The Naïve Bayesian model used is optimized for sparse datasets, implemented as described in Xia et al. [5].
The Bayesian model generates a score that can be used to rank compounds that have not been screened yet.
The machine learning model provides a set of parameters as well as the ROC plot so you can gauge its effectiveness. Each molecule receives a relative score, applicability number, and maximum similarity number.





The model automatically scores all compounds in the selected project. It can subsequently be shared with other projects to score more molecules.



CONCLUSION

We demonstrated that the discovery of antibiotic and infectious diseases drug candidates can be potentiated via an effective scientific-community based collaborative research. Aditeren has been identified as a potential good candidate for crossing Gram-negative bacteria membranes to the Cytosol.
These results are based on data curated and publicly available freely in the CDD Vault public access section[3].

Acknowledgments

The authors would like to thank: Dr Lynn Silver, who has made possible most of this work