excelra

GLAMOROUS AI

Comparative Analysis of Training Machine Learning Algorithms on Different Data Scales

Prediction of Solubility & LogD

WHITEPAPER

Authors

Noor Shaker

Mohamed Abou-Zleikha

Sergio Pascual

Anil Kumar Manchala

Vani Gudimella Tirumala

Norman Azoulay

Background

Poor drug solubility is one of the main obstacles in drug discovery and development and is strongly related to the choice of target explored. (Bergstrom et al., 2016). Solubility is critical for absorption and acceptable solubility in the intestinal fluid is required to achieve sufficiently high drug blood concentrations to obtain a therapeutic effect when systemic effects are warranted. The solubility of a compound affects its absorption, distribution, metabolism, excretion and toxicity (ADMET) profile. Only drug candidates whose ADMET properties are of sufficient quality can be further developed.

Introduction

SAR Databases

Machine learning methods are reshaping research on properties of molecules. The better the datasets that are used to train and test these methods, the more robust the results are expected to be. In this white paper, we compare two such SAR databases: GOSTAR, which is the largest Structure Activity Relationship (SAR) database for drug discovery, and MoleculeNet, which comprises multiple public datasets, establishes metrics for evaluation, and offers open-source implementations of multiple previously proposed algorithms for evaluating compounds using machine learning.

The MoleculeNet collection includes over 700,000 compounds tested on a range of different properties. The benchmark also tests performances of various machine learning models with different featurizations on the datasets and results reported in AUC-ROC, AUC-PRC, RMSE and MAE scores.

GOSTAR (Global Online Structure Activity Relationship) provides a 360^{o} view of over 8 million small molecule discovery compounds and close to 50,000 preclinical/ clinical candidates, and approved drugs. Content in GOSTAR is meticulously curated manually from various published data sources with information on chemical structures and their biological properties that includes binding, in-vitro, in-vivo, ADME, Tox and physicochemical properties.

Datasets (LogD and Solubility)

Our study, focused on solubility, started with over 9,000 molecules from the GOSTAR dataset, approximately 6,800 of which were unique. After data cleaning and preprocessing, the total set of valid records contains 21360 for the GOSTAR data and 4200 for MoleculeNet data. These are the dataset we used for the benchmark in this work.

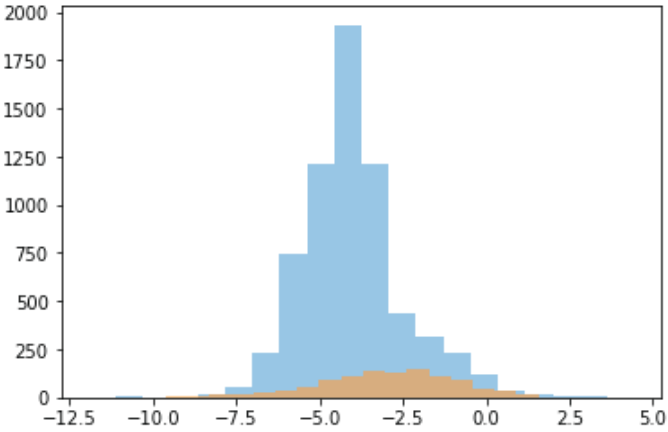


Figure 1. Distribution of Solubility Data from GOSTAR and MoleculeNet

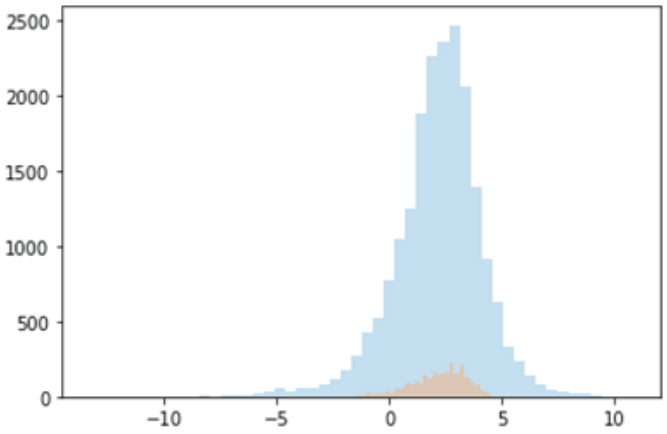


Figure 2. Distribution of LogD Data from GOSTAR and MoleculeNet

Methodology

Glamorous[AI] used its flagship platform, RosalindAI, to build, train and evaluate a zoo of machine learning models that predict solubility and logD. Rosalind[AI] is an end-to-end platform for managing molecular data and developing cutting-edge ML pipelines in a scalable and reproducible manner without the need for coding skills. Rosalind[AI] automatically ingests molecular data, and executes pipelines of automated processes and modeling that include data cleaning and preprocessing, featurization, model initiation, hyper-parameter optimization and model benchmarking.

For the purpose of this study, Rosalind[AI] is used to clean and prepare the data, develop and train the models and report the results. We used random forest for benchmarking as it is the golden standard. For best performance, we used a zoo of models that includes different types of deep learning models trained in a supervised manner on different molecular featurisations (SMILES, Graph, RDKit, Morgan, etc). Rosalind[AI] allows testing of a large number of modeling architecture and hyperparameter optimization strategies to ensure convergence to the best model.

Results & Discussion

Glamorous^{AI} Result Set

Solubility Prediction – training and benchmarking procedure – with GOSTAR data, glamorous is 9.5% better than random forest. With public data, glamorous is 14% better than random forest.

Data Size	Model	LogD (RMSE, lower is better)	Solubility
21,360 (Gostar) 66,05 (Sol)	**Rosalind^{AI}**	**0.83**	**0.85**
	Random Forest	1.5	0.94
4,200 (Public) 1,085 (Sol)	**Rosalind^{AI}**	**0.55**	**0.56**
	Random Forest	0.98	0.7

We ran a comparison of 6,605 GOSTAR molecules versus 1,085 from MoleculeNet. The bottom line was that Glamorous^{AI} models trained on the GOSTAR database, are generally 2x or 92% better at predicting solubility than models trained on MoleculeNet's dataset.

Many off-target liabilities, such as plasma protein binding (especially albumin), hERG, CYP interactions, and transporters, have strong correlations with lipophilicity, and a number of studies have linked high logD to the likelihood of compounds failing in development due to poor ADMET (absorption, distri bution, metabolism, excretion and toxicity) character istics. The majority of known drugs contain ionizable groups and are likely to be charged at physiological pH. The distribution constant, LogD, is therefore a better descriptor of a molecule's lipophilicity than logP. LogD is thus pH dependent, so the pH at which the logD was measured must be specified. As the physiological pH of blood serum is 7.4, logD7.4 is of particular interest.

2x
BETTER AT
Predicting Solubility

Conclusion

It's clear from this study that GOSTAR's proprietary data set combined with rigorous processes for data cleaning and large scale ML training and development deliver far more robust and actionable results. The larger number of data points, intensive curation, and available trouble-shooting for cleaning up data make this a much better user experience. Side-to-side studies of Glamorous^{AI} application to GOSTAR and public data for solubility analysis shows that Glamorous^{AI} provides substantially better results with either dataset.

About

GlamorousAI

GlamorousAI is a London-based startup aiming at bringing drugs to every target. It's flagship platform, RosalindAI, democratizes access to cutting-edge AI and specializes in solving data challenges that makes the majority of disease associated targets intractable by existing computational technologies. By solving small, sparse and noisy data problems in drug discovery, RosalindAI demonstrates higher hit rate, better model accuracy, robustness and significant reduction in lab experiments.

30%
Hit Rate

8%
More Accurate

50%
Less Data

3x
More Targets

GOSTAR

From target profiling to hit identification and lead optimization, GOSTAR is the perfect resource for the medicinal and computational chemists capturing granular assay data across chemical, biological, pharmacological and therapeutic dimensions. Further, Excelra provides custom curation support for bespoke client needs, data preparation for AI/ML modeling and secure curation on the cloud of proprietary client data.

28M
SAR Activities

8M
Chemical Structures

76K
Targets

36K
End Points

References

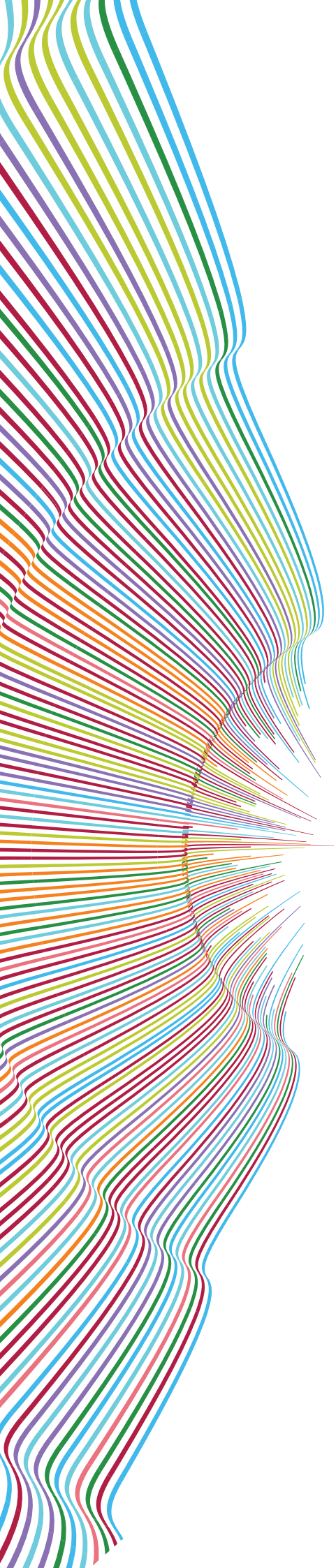
- Chen H, Engkvist O, Wang Y et al (2018) The rise of deep learning in drug discovery. Drug Discov Today 23. https://www.sciencedirect.com/science/article/pii/S1359644617303598

- Bhhatarai B, Walters WP, Hop C et al (2019) Opportunities and challenges using artificial intelligence in ADME/Tox. Nat Mater 18. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6594826/

- Tang B, Kramer ST, Fang M et al (2020) A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. J Cheminform 12. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-0414-z

- Jiang D, Wu Z, Hsieh C-Y, et al (2021) Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform 13. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00479-8#Sec24

- Wu Z, Ramsundar B, Feinberg E N, et al (2017) MoleculeNet: A benchmark for molecular machine learning. arXiv.org MoleculeNet: a benchmark for molecular machine learning



excelra

Transforming Life Science Data into Actionable Insights

HYDERABAD • BOSTON • UTRECHT

Reach us at: marcom@excelra.com

www.excelra.com

GLAMOROUS AI

Bringing Drugs to Every Target

LONDON

Reach us at: contact@glamorous.ai

www.glamorous.ai