American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Research Article

AI-Based Predictions of Molecular Target Activity from Blind Chemical Structures

Ian Jenkins[1], Vaishnavi Narayan[1], Jayson Uffens[1], Eric J Mathur[1], Saman Mirzaei[1], Krista Casazza[1] and Jonathan RT Lakey[1,2,3*]

[1]GATC Health Corp., 2030 Main Street, Suite 660, Irvine, CA

[2]Departments of Surgery and Biomedical Engineering, University of California Irvine, Irvine, CA

[3]Department of Cardiovascular Research, Faculty of Medicine, University of West Virginia, Morgantown, WV

***Corresponding author:** Jonathan R T Lakey, Department of Surgery, University of California Irvine, Irvine, California, USA, 92868, California, USA.

To Cite This Article: *Ian Jenkins, Vaishnavi Narayan, Jayson Uffens, Eric J Mathur, Saman Mirzaei, Krista Casazza and Jonathan RT Lakey*. AI-Based Predictions of Molecular Target Activity from Blind Chemical Structures. Am J Biomed Sci & Res. 2024 23(1) AJBSR.MS.ID.003034,* DOI: 10.34297/AJBSR.2024.23.003034

Received: June 17, 2024; **Published:** June 21, 2024

Keywords: Machine learning, Artificial intelligence, Multi-therapeutic interventions augmented intelligence (AI), Drug discovery

Introduction

Accurately identifying protein-small molecule interactions is the cornerstone of modern drug development. These molecular interactions have enabled elucidation of the underlying biological processes, resulting in effective in silico therapeutic design. As the primary functional molecules in cells, proteins play critical roles in virtually all biological activities, from catalyzing metabolic reactions to transmitting signals within and between cells. Therefore, understanding how proteins interact with exogenous molecules, including small drug-like compounds, is essential for elucidating function and rational design of new drugs and employment of a rational poly-pharmaceutical approach.

Identifying the interactions between proteins and small molecule compounds can reveal new therapeutic targets. Many infectious diseases and cancers arise from aberrant protein interactions. By understanding these interactions, researchers can identify critical nodes within cellular pathways that, when modulated by a drug, may correct the underlying disease mechanism [1].

The inherent complexity of biological systems makes it challenging to predict how a drug will interact within the human body. Traditional methods often failed to accurately model these inter actions, leading to failure of successful endpoints in clinical trials. Although the advent of advanced computational technologies has attenuated failure rate for drug development, it remains high. The reasons for these failures are highly variable, but most commonly include issues related to efficacy, safety, and pharmacokinetics. Approximately 90% of drug candidates fail during the preclinical phase due to unacceptable toxicity or lack of efficacy. In Phase 1, approximately 30% of the remaining drugs fail. In Phase 2, 67% of these remaining drugs fail. In Phase 3, another 40% of drugs that passed Phase 2 fail due to unforeseen side effects or inability to reach the defined endpoint [2,3].

Modern machine learning tools in computational biology, exemplified by the GATC Health multi-omic therapeutics platform, have revolutionized the ability to predict protein-molecule interactions with high accuracy. These tools integrate deep learning with vast amounts of structural and interaction data, enabling researchers to model complex molecular systems that were previously challenging to study experimentally. This integration of computational predictions with experimental data accelerates the drug discovery pipeline, from target identification to lead optimization.

As a validation study, a blind challenge was conducted to assess the performance of these advanced, predictive capabilities of the GATC platform in a fair and unbiased manner and provide a standardized benchmark against which different AI platforms can be compared. Using the same dataset and evaluation criteria, GATC obtained an objective measure of the strengths and weaknesses of various algorithms within the platform. The blind challenge serves as a continuous checkpoint to assess the progress in the platform over time. By comparing the performance of the GATC platform, advancements can be tracked and areas that require further improvement can be identified.

Methods

University of California, Irvine, (UCI) Department of Pharmaceutical Sciences lab partnered with GATC to administer the evaluation challenge ensure a rigorous, fair, and unbiased evaluation of the algorithms through a blind challenge. UCI served as the neutral party to handle the test data and evaluation process. GATC did not have access to the test set.

Data Curation

The identity and sources of GATC's training data for models used in this study were fully disclosed to the UCI Lab. This ensured that data curated by the UCI Lab for the challenge was known to be outside of GATC's training data. GATC also provided the identity of available biological targets that the GATC platform could predict.

The UCI Lab then curated (from the literature and/or other sources) data outside GATC's training data concerning binding of molecules to these targets as the basis for running a blind challenge.

This allowed for focus on chemical compounds that did not overlap with those in GATC's training data, but which have been experimentally tested for binding to targets the GATC platform covers, as provided by GATC.

For the purposes of data curation and the blind challenge, a compound was considered "active" if it had a measured potency (e.g. IC50 or Kd or similar) better than 1 micromolar and "inactive" if it has a measured potency worse than 10 micromolar.

The UCI Lab then collected a relatively balanced set of data which included a roughly equal number of active and inactive compounds, with the total number being over 1000 and less than 10,000 for each challenge.

The resulting datasets were then prepared by the UCI Lab for the blind challenges in the form of a single dataset for each challenge with compound IDs and SMILES strings, called the "prediction dataset".

The UCI Lab separately stored and held in secrecy an "answer key" with assay results/activities and other metadata including target UNIPROT ID, binary activity of true or false, and measured activity, as well as data source, among others for the challenge.

Blind Challenges

Following curation of data for the blind challenges, the UCI Lab provided GATC with an example file illustrating the dataset format, to ensure GATC could parse the data properly.

Once the data format transfer was agreed upon, the UCI Lab coordinated with GATC and set up a synchronous meeting during which the "prediction dataset" will be provided to GATC and GATC will synchronously generate and return predictions to the UCI Lab, in a .csv format, including compound ID, SMILES, binary activity, and UNIPROT ID of target.

Following receipt of this data and verification that the data can be parsed properly, the UCI Lab returned the answer key for GATC's use, concluding the blind challenge.

Key measures for each challenge were determined to be the True Positive Rate or Sensitivity, and the True Negative Rate or Specificity of the prediction data set as compared to the original, blind challenge dataset for activity vs inactivity for each combination of molecule and target.

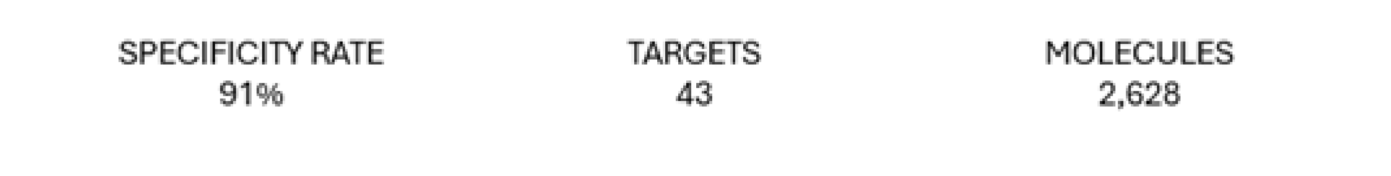
As each molecule presented in the challenge dataset may not have corresponding activity for each target available on the GATC platform, predictions were made for every molecule/target combination and included in the prediction dataset.

Only predictions with matching molecule/target combinations in the challenge dataset were included in results calculations.

Results

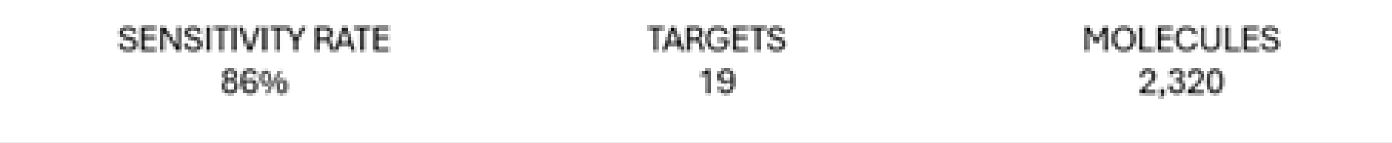
The test molecules were confirmed to be previously unknown as GATC to curate a fair blind challenge. For each of the challenges, the GATC platform processed the molecule data within hours and returned its predicted results. Only after receiving these results did the UCI Lab release the target activation data for the challenge molecules to GATC.

The results below are taken from a comparison of the activation data curated and provided by the UCI lab in each challenge dataset and the blind predictions made by the GATC platform and included in the prediction dataset for each of the (challenge 1,2).



SPECIFICITY RATE	TARGETS	MOLECULES
91%	43	2,628

Challenge 1: Specificity Screening - predictions by AI to screen for negative activation on biological targets. Key in assessing the safety and side-effects of a drug molecule.



SENSITIVITY RATE	TARGETS	MOLECULES
86%	19	2,320

Challenge 2: Sensitivity Screening - predictions by AI to screen for positive activation on biological targets. Key in assessing efficacy of a drug molecule and secondary assessment of safety and side-effects.

Discussion

The performance of the GATC platform on identifying activity both on the safety and efficacy of drug candidates was sufficient to warrant use in risk assessment of drug candidates within the defined areas of work and on the most deleterious associated risk activities. There is relatively little data available to use in blind challenges that is not already used to train the GATC platform as the training has been extensive. In order to provide a good assessment of the platform's capabilities, the blind data used for these challenges was obtained from broader sources of public data by the University of California, Irvine. The data used for Challenge 2 includes a smaller number of targets, as defined by the safety assessment on higher-risk associations. This second set of challenge data has also been divided and GATC has held a portion in reserve to support ongoing research and platform improvements.

The metrics reported are based on data filtered within categories of commercial interest based on prospective work in risk assessment within. The data filters used in this challenge do not necessarily encompass the entirety of possible interactions within the human body or interactions outside of the scope of commercial interest. Future work will include broader and more diverse targets. GATC reports that the targets were selected for a particular research focus, and the proprietary methods used by the GATC System to achieve the reported specificity and sensitivity are not standard and may be considered trade secrets.

Ideally, a future challenge might either add such optimizations to other comparator platforms or measure both sensitivity and selectivity based on older standards of computational performance to avoid any potential bias in analysis across platforms which may not have such optimizations as the GATC platform. As such, the blind challenge is instrumental in fostering transparency, objectivity, and progress in the field of artificial intelligence by providing a standardized platform for evaluation and comparison.

Acknowledgments

The author acknowledges the support for this study from GATC Health Inc. and our collaboration with the University of California Irvine.

References

1. Ah Ram Kim, Yanhui Hu, Aram Comjean, Jonathan Rodiger, Stephanie E Mohr, et al (2024) Enhanced Protein-Protein Interaction Discovery via AlphaFold-Multimer. bioRxiv 02(19) 580970.
2. Wong CH, Siah KW and Lo AW (2019) Estimation of clinical trial success rates and related parameters. Biostatistics 20(2): 273-286.
3. Callaway Ewen (2022) "What's next for AlphaFold and the AI Protein-Folding Revolution." Nature 604(7905): 234-238.