CYCLICA

VALIDATION STUDY

MatchMaker Validation on newly published protein-ligand interaction dataset

Proteome screening, powered by Cyclica's MatchMaker deep learning engine, predicts primary protein targets of newly published ligands

Overview

- MatchMaker, Cyclica's biophysics-based deep learning engine for predicting drug-target interactions, was applied to the identification of targets for a dataset of recently published protein-ligand interactions.
- MatchMaker ranked 55% of the targets amongst the top 1% of proteins predicted to bind their respective ligands.
- Comparative analysis of performance with the four most recent releases of MatchMaker revealed that accuracy improved over time.

In silico Proteome Screening of newly discovered ligands

MatchMaker is Cyclica's proprietary deep learning engine for predicting drug-target interactions (DTIs), trained on millions of DTI pairs across the human proteome, encompassing over 8,500 proteins to date. To evaluate the performance of MatchMaker versions released over the course of 2020, we applied MatchMaker to the prediction of drug targets for 97 recently published protein-ligand pairs. These pairs were retrieved from 97 peer reviewed articles published in medicinal chemistry journals between June and December 2020. Those which confirmed a direct interaction between a lead compound and a primary target protein through either biophysical or biochemical assays were retained, yielding a total of 97 unique compounds that interacted with 81 unique proteins. One of the compounds was a dual-acting ligand with two target proteins, yielding a total of 98 unique protein-ligand interactions in our datasets. The dataset encompasses a wide diversity of target classes, with GPCRs, kinases and enzymes representing the largest proportion of protein families.

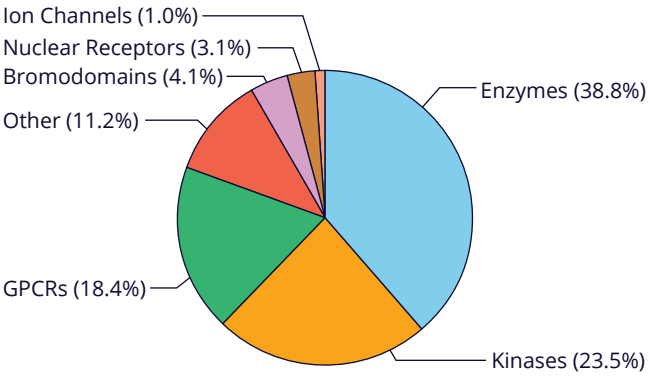

Figure 1. Protein family distribution of the 81 unique proteins associated with the 97 lead compounds retrieved from recent medicinal chemistry journals spanning June to December 2020.

MatchMaker-based Proteome Screening generates polypharmacological profiles of compounds against the structurally-characterized human proteome. Protein binding sites are ranked based on their likelihood of interaction with a query compound. Over the course of 2020 we released four MatchMaker models (Q1, Q2, Q3, Q4); in this study we executed proteome screens of the 97 lead compounds using each MatchMaker model, and compared the ranks of each target protein across all four models. For each of the 81 proteins, we compared the ligand from the test set to the ligands present in the training set for the particular protein involved in the protein-ligand pair. The median tanimoto similarity between the tested ligand and those present in the training set that are associated with the target protein ranged from 0.4 to 0.5 across the four models (Morgan3, 1024-bit, no chirality fingerprint) (**Table 1**). Notably, some interactions with very low tanimoto similarity (<0.3) were ranked above the 95th rank percentile, demonstrating the power of our predictive engine.

	Q1	Q2 & Q3	Q4
Mean	0.43	0.48	0.53
Min.	0.24	0.26	0.26
Max.	0.83	1	1
Median	0.40	0.46	0.49
1st Quartile	0.33	0.36	0.39
3rd Quartile	0.48	0.57	0.61

Table 1. Summary statistics of tanimoto similarity between ligands in the testing set and ones present in MatchMaker's training sets associated with the particular target protein involved in the protein-ligand interaction.

MatchMaker's predictive accuracy increased throughout 2020

In this study, we observed an improvement across models on the 98 newly discovered protein-ligand interactions. MatchMaker ranked 30%, 47%, 45% and 55% of the targets amongst the top 1% of proteins predicted to bind their respective ligands in Q1, Q2, Q3 and Q4 respectively (**Figure 2** and **Figure 3**).

We hypothesized that this improvement is due to the increased number of interactions present in the Q4 training set, where 8 interactions with a tanimoto similarity of 1 were present. However, removing exact interactions present in Q2, Q3 or Q4 from our analysis did not alter the observed trend (**Figure 4**) and 52% of known targets still ranked in the top 1% after excluding them. Moreover, the number of target proteins ranking in the top 50 nearly doubled from Q1 to Q4 (24% to 48%) (**Figure 5**). When removing protein-ligand interactions that were present in Q4's training set (tanimoto similarity of 1), the number of target proteins ranking in the top 50 still increased from 24% in Q1 to 44% in Q4. These improvements can in part be attributed to a significant enlargement of the training dataset (4,371,884 new unique drug-target-interactions added from GOSTAR database in Q4), as well as to the implementation of transfer-learning algorithms, both of which serve to improve the robustness and, consequently, generalizability of MatchMaker.

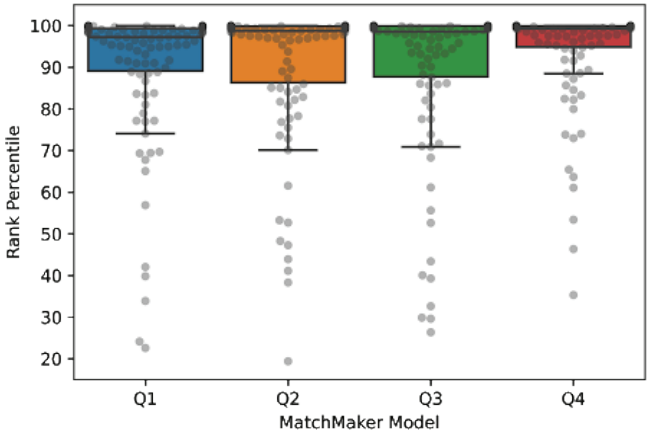

Figure 2. 98 protein-ligand interactions' rank percentiles across MatchMaker models. MatchMaker's prediction accuracy increased with new releases.

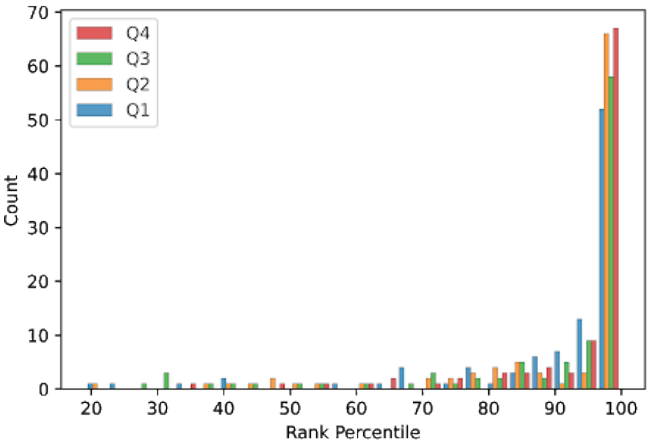

Figure 3. Distribution of MatchMaker models' rank percentiles for the protein-ligand interactions.

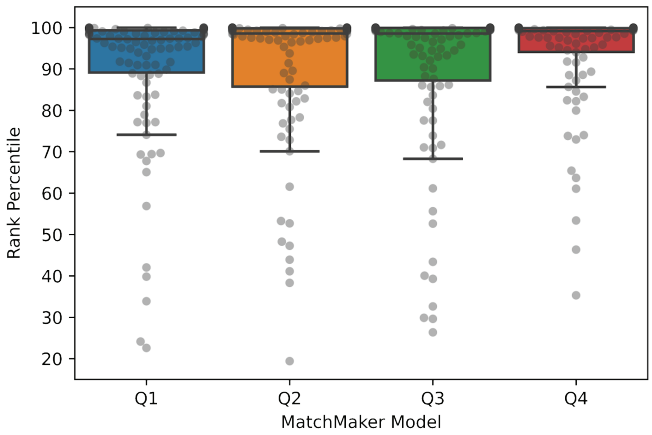

Figure 4. Rank percentiles of protein-ligand interactions upon removing those present in Q4's training set (tanimoto similarity of 1).

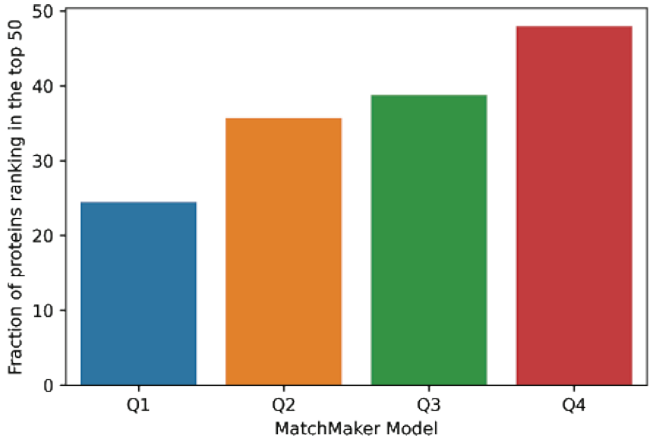

Figure 5. Number of target proteins ranking in the top 50 against the human proteome (>8500 proteins).

The fluctuations in frequency of rank percentiles shown in **Figure 4** is not unexpected. MatchMaker releases have been tested across multiple protein classes and not on the specific set of protein targets in this study. The monotonic increase in performance we achieved for MatchMaker across Q1-Q4 models, using our testing strategies, may not be representative of every possible set of proteins. This fits expectations since machine learning models need to be tested using more comprehensive testing strategies rather than specific protein sets, except if the model is directly developed for that set.

Summary

With MatchMaker powered Proteome Screening, we retrospectively validated 98 recently disclosed protein-ligand interactions. MatchMaker ranked 55% of the targets amongst the top 1% of proteins predicted to bind their respective ligands. Within this real-world example, our models showed improved predictive performance each release. MatchMaker's predictions are accurate and generalizable, making them useful in the experimental process by offering insights into the polypharmacology profiles of the query compounds.