CYCLICA

VALIDATION STUDY

Comparison of MatchMaker to DeepConv-DTI reveals superior performance and compute efficiency for predicting drug-target interactions

Cyclica's MatchMaker deep learning engine excels at dataless protein targets

Overview

- MatchMaker and DeepConv-DTI compared head-to-head on large-scale Drug-Target Interaction (DTI) datasets with >1,000,000 active protein-ligand interactions, with application-relevant cross-validation tests.
- MatchMaker demonstrates substantial prediction performance gains on novel targets and modest gains on novel compounds.
- MatchMaker is orders of magnitude faster, using >400-fold less CPU time for training and >100-fold less for prediction.

Introduction: Predicting drug-target interaction

Modeling and predicting drug-target interaction (DTIs) are the most common tasks performed by practitioners of computer-aided drug design, with applications ranging from target identification to design of new chemical entities (NCEs). Traditionally these methods fall into one of two classes: target-based, such as docking, and ligand-based, such as QSAR models or one-shot learning approaches. However, over the last several years a new class of machine learning models for DTI prediction has emerged; known as *'DTI models'* or *'DTI predictions'*, these models learn binding patterns from large-scale bioactivity datasets[1] by combining features derived from both protein targets and small molecules to distinguish binding from non-binding protein-ligand pairs (Figure 1). To reduce model bias, ideally these datasets should span a diversity of protein classes. Further, protein representations that capture salient properties related to ligand binding have the greatest potential to generalize among protein systems. As

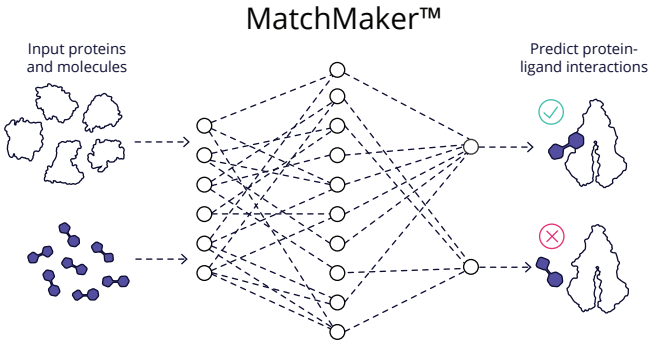
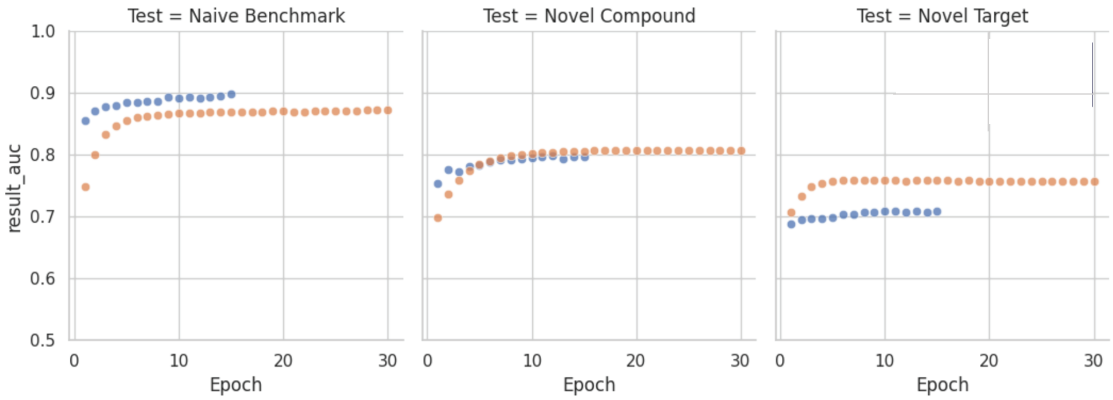


Figure 1. A schematic representation of DTI prediction models that combine input representations of proteins and small molecules to subsequently evaluate interaction predictions.

compared to target-specific ligand- and structure-based models DTI models are much better suited for applications such as multi-target screening, target ID, or addressing novel protein targets. Moreover, structure-based approaches, including machine-learning based models[2], are inherently dependent on accurate prediction of binding geometry, introducing predictive volatility and high compute costs.

MatchMaker, Cyclica's DTI prediction model, is trained on ~1.5M human bioactivities, sourced from public and private sources. MatchMaker differs from literature-reported DTI models by introducing structural context to its protein representations. We accomplish this through systematic mapping of all DTI pairs onto presumed binding sites of 3D protein structures - both experimentally-determined and modelled - using a combination of pocket detection, homology mapping and structural superposition methodologies. Models are then trained using the assumed pocket features to represent each protein, while simultaneously addressing uncertainties in data quality and pocket mapping with *Filtered Transfer Learning*[3].



Final MatchMaker AUC: 0.872
Final DeepconvDTI AUC: 0.899

Final MatchMaker AUC: 0.808
Final DeepconvDTI AUC: 0.798

Final MatchMaker AUC: 0.757
Final DeepconvDTI AUC: 0.710

Figure 2. Validation results demonstrating superior predictive performance by MatchMaker on *Novel Compound* and *Novel Target* tests.

In this study we compare the performance of MatchMaker for predicting drug-target interactions to DeepConv-DTI, an open-source, deep-learning based model that represents proteins using a primary sequence based convolutional neural network[4]. We demonstrate superior performance of MatchMaker on extrapolation to novel protein-ligand systems, i.e. where no data related to the molecule or target was used to train the model and comparable performance in other scenarios. Further, from a compute perspective both training and evaluation time are orders of magnitude faster than DeepConv-DTI, enabling considerable scaling at a fraction of the compute cost.

Methods

We designed three separate training experiments to evaluate ligand-dependent and protein-dependent generalizability, described below:

- Naive Benchmark Test: Excludes all bioactivities for ligands present in the test set.
- Novel Compound Test: Excludes all bioactivities for ligands sharing > 0.40 Tanimoto similarity to any test set ligand.
- Novel Target Test: Excludes all bioactivities for ligand or proteins present in the test set.

DeepConv-DTI was trained with the optimal parameters described in its source publication, including a 1:1 simulated negative training data ratio and 15 training epochs. MatchMaker was trained in accordance with its internal 2021Q1 release protocol, which uses a 1:19 simulated negative training ratio and 30 training epochs. A static test set was created from 300 proteins present in the bioactivities dataset, selected to maximize class diversity and avoid bias towards high-data targets. The test set contains 5527 total DTI measurements involving the 300 proteins and 5050 distinct ligands. Three separate training sets were derived from Cyclica's 1.5m bioactivities dataset by applying separate exclusion criteria, based on similarity of ligands or proteins to the static test set. Tanimoto similarities are based on RD-Kit's Morgan fingerprint (radius 3, 1024 bits).

Results

Cross-validation results for all three test scenarios are summarized in Figure 2. The Naive Benchmark Test represents a baseline scenario, where the model is applied to protein systems with known bioactivities and compounds similar to known bioactives. On this Naive Benchmark test, DeepConv-DTI outperforms MatchMaker. We attribute the increased capacity for data memorization to DeepConv-DTIs more complex network architecture and different priorities in model development. Notably, this metric for model performance is subject to *compound series bias*[5] as significant chemical redundancies exist within compound series for a given drug discovery program. This redundancy often results in high similarity, both in terms of structure and activity, between training and test datasets.

The Novel Compound test represents an application scenario where the test molecules represent new molecular scaffolds with no reported bioactivities. The strict threshold excludes ~32% of all ligands from the training dataset and is likely to contain the analogs or/and lead fragments to known actives. On this ligand-dependent generalizability test, MatchMaker demonstrated a modest improvement in predictive performance over DeepConv-DTI.

The Novel Target test is representative of dataless target applications. Here, MatchMaker demonstrates

	MatchMaker	DeepConvDTI
Benchmarking parameters	30 Epochs 19:1 Simulated Negatives	15 Epochs 1:1 Simulated Negatives
Model Training Time	264 CPU*Hours	5856 CPU*Hours
Model Evaluation Time	789 pairs per CPU*S	7.17 pairs per CPU*S

Table 1. MatchMaker models train and evaluate DTI pairs with high CPU efficiency, providing opportunities to scale model applications and expand training datasets.

clear performance gains relative to DeepConv-DTI. We attribute MatchMaker's target generalizability to the structurally-augmented representation of proteins. Superior protein generalizability positions MatchMaker as the ideal platform to discover hits on new targets or to perform proteome screens aimed at phenotypic deconvolution.

Last, we report significant differences in the computing needs of both algorithms (Table 1). Computational efficiency of DTI prediction engines can impose constraints on model development and model applications. Despite training twice as many epochs and using 19x as many simulated negatives, MatchMaker models still train on a fraction of the computational resources required to build DeepConv-DTI models. When adjusting for simulated negative ratios and epoch counts, MatchMaker demonstrates a 443x computational efficiency in model training. Cyclica's *Machine Learning Team* benefits from efficient model training by continually performing 100s of training experiments and optimizations to improve performance on a quarterly basis. Moreover, MatchMaker evaluates DTI pairs 110x faster than DeepConv-DTI. Faster model evaluation allows Cyclica to exhaustively screen virtual compound libraries exceeding 1 billion molecules or proteome cross-screens (~9k proteins) on compound libraries exceeding 1 million molecules.

Concluding remarks

Cyclica's platform combines the best of both of the structure-based and ligand-based approaches in a model named MatchMaker. This combination gives MatchMaker the ability to generalize from first principles in biophysics as well as from patterns in the experimental data. Generalization is further amplified by the use of proteome-wide training data. This ability to generalize enables the prediction of binding to proteins with no known binders, and even the prediction of binding to completely novel pockets, e.g. for the design of compounds with allosteric activity. This approach is unique in the industry and explains the high rate of success we have seen discovering small molecules for our partners, some of them for very difficult targets.

References

1. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W. & Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **24**, i232–40 (2008).
2. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein-Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
3. Madani Tonekaboni, S. A. *et al*. Learning across label confidence distributions using Filtered Transfer Learning. in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* 1117–1123 (2020). doi:10.1109/ICMLA51294.2020.00180.
4. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol*. **15**, e1007129 (2019).
5. Mayr, A. et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **9**, 5441–5451 (2018).