Improving Hit-finding: Multilabel Neural Architecture with DEL

Kehang Han[*]
Google Research
kehanghan@google.com

Steven Kearnes [†]
Relay Therapeutics
skearnes@relaytx.com

Jin Xu
Google Research
jix@google.com

Wen Torng
Google Research
wtorng@google.com

JW Feng
Google Research
jwfeng@google.com

Abstract

DNA-encoded library (DEL) data, often with millions of data points, enables large deep learning models to make real contributions in the drug discovery process (e.g., hit-finding). One recent success method of modeling DEL data [1], GCNN [2] multi-class model, requires domain experts to create mutually exclusive classification labels from multiple selection readouts of DEL data, which is not always an optimal formulation. In this work, we designed a multi-label architecture that directly models each selection data to eliminate the corresponding dependency on human expertise. We selected effective choices for key modeling components such as label reduction scheme from *in silico* evaluation. To assess its performance in real-world drug discovery settings, we further carried out prospective wet lab testing where the multi-label model shows consistent improvement in hit-rate (percentage of hits in a proposed molecule list) over the current state-of-the-art multi-class model.

1 Introduction

One of the major barriers in applying machine learning, especially deep learning, to drug discovery problems is limited data. The recent advent of DNA-encoded libraries (DELs) with their massive data size opens many new opportunities [3] [4] [5]. Training large deep learning molecular property models becomes possible, and applying them to key stages (e.g., hit-finding, hit to lead) of the drug discovery process has shown great promises [1].

The readout of an experimental DEL selection is DNA sequence counts [3], which are commonly aggregated into disynthon representations to calculate enrichment scores with good signal-to-noise properties [1]. Besides the primary DEL selection experiment (measuring binding signal called "Target Enr" when protein target is present, i.e., on-target binding experiment), control selection (measuring binding signal called "NTC Enr" when no protein target is present, i.e., No-Target Control binding experiment) and additional counter-selections could be run to further reduce noise, for example, inclusion of a known competitive inhibitor (measuring binding signal called "Competitor Enr" when protein target is competitively inhibited) [1].

In McCloskey et al.'s work [1], data from those multiple DEL selections are combined to create mutually exclusive labels corresponding to different experimental outcomes. These labels represent logical combinations (e.g. AND/OR) of enrichment scores from the multiple experimental conditions

[*]Work done as an AI resident.
[†]Work done in Google Research

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

as well as external side information such as hit frequency across targets denoted as "Target ratio" (to identify promiscuous compounds that are unlikely to have specific interactions with the target). The label derivation can be often summarized in a decision tree, which we call a *column scheme*. See an example in Figure 1.

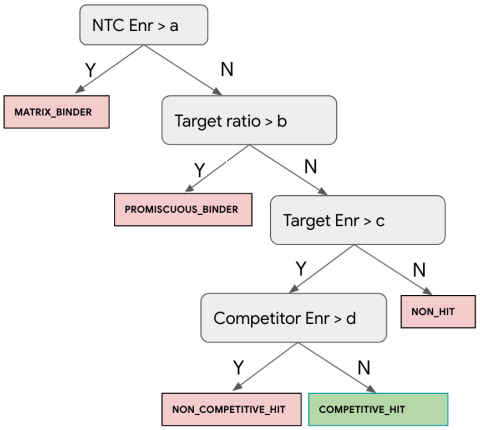


Figure 1: Example column scheme: derivation of class labels that are needed in GCNN multiclass models. "NTC Enr" stands for molecule's enrichment score in No-Target Control binding experiment, "Target Enr" stands for a molecule's enrichment score in on-target binding experiment, "Competitor Enr" stands for a molecule's enrichment score in on-target binding experiment but the protein target's desired binding pocket has been occupied by other molecules already, and "Target ratio" is statistic quantifying how frequently a molecule binds to any protein target

A classification model can then be readily built with the derived classes. One major problem with this approach is that in reality a molecule being categorized as one class can still be of another class (e.g., a promiscuous compound can still be a target competitive hit). Additionally, using labels derived from multiple experiments, each with their own criteria and thresholds, requires the model to learn a complicated latent structure for the human-crafted labels (fixed at training time, see Figure 1). Our key insight is that the separate DEL experimental outcomes can be treated independently. In this way, the label naturally matches with the physical meaning of each selection experiment—the resulting model makes a prediction for each experimental outcome and these predictions can be flexibly combined as needed in downstream applications. We summarize our contributions in this paper as follows:

- we design and implement key components of a multi-label neural architecture that models DEL data more naturally (Figure 2),
- we conduct *in silico* retrospective testing experiments and demonstrate this philosophy of modeling each DEL selection experiment independently can bring better performance than the multi-class architecture,
- to examine performance in real drug-discovery settings, we carry out wet lab prospective testing experiments [6] for two protein targets, which showed superior performance of multi-label models in hit-finding applications.

2 Results

In order to compare hit-finding performance with the current state-of-the-art GCNN multi-class model [1], we selected two relatively hard protein targets: Tyrosine-protein kinase (c-KIT) and Estrogen Receptor Alpha (ERa). In [1], c-KIT and ERa show relatively low hit rates (9.7% and 18.8% respectively at concentration of 10uM), giving this study more room to improve. We trained two types of graph neural networks to enable direct comparison: GCNN multi-class model used in [1] and GCNN multi-label model developed in this study (see Figure 2).

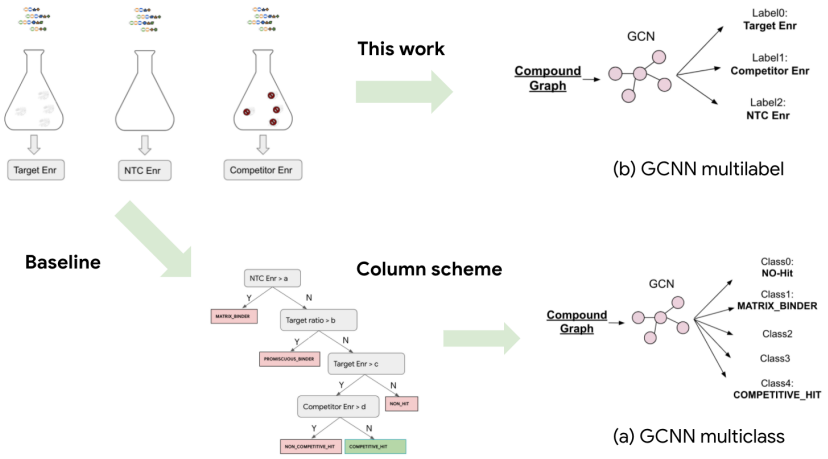


Figure 2: Modeling setup: the baseline is a GCNN multi-class model which requires a column scheme designed by human expertise; this work reports a GCNN multi-label model that learns from each selection individually.

Table 1: *in silico* retrospective evaluation against two c-KIT inhibitor datasets and five internal ERa inhibitor datasets for GCNN multiclass model and two multilabel models with variation on label reduction scheme: (a) simply using target enrichment label as final reduced label, i.e., Reduced-label = Target-Enr-label, (b) using Reduced-label = Target-Enr-label − NTC-Enr-label − Competitor-Enr-label

actives@100 (↑)	c-KIT-test0	c-KIT-test1
GCNN multiclass	27	59
GCNN multilabel (a)	**28**	**61**
GCNN multilabel (b)	**29**	57

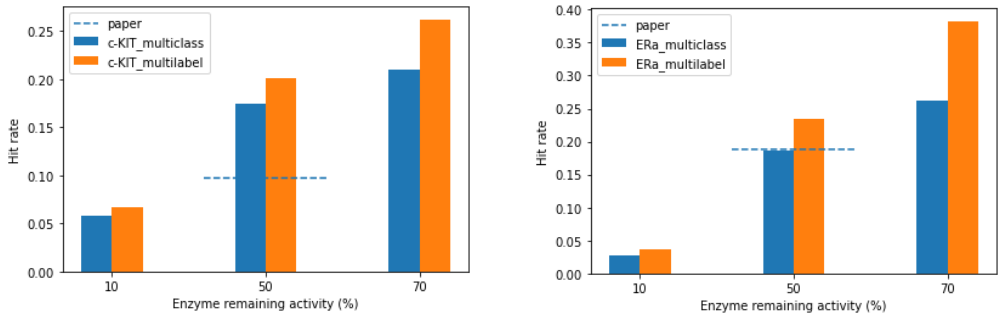
actives@100 (↑)	ERa-test0	ERa-test1	ERa-test2	ERa-test3	ERa-test4
GCNN multiclass	99	77	42	23	**13**
GCNN multilabel (a)	**100**	**79**	**43**	**26**	11
GCNN multilabel (b)	**100**	**79**	**45**	**26**	12

2.1 *In silico* evaluation

As a first step, we carried out *in silico* retrospective evaluation of our models against internal inhibitor datasets. Specifically, we have five test datasets for ERa and two for c-KIT. We monitor a metric that is closely related to hit-rate: the number of active molecules among the top 100 highest scoring molecules in the test set (actives@100). Table 1 shows the results. Despite the slight improvement, the GCNN multi-label models outperforms its multi-class counterpart for most of the test sets (6 out of 7).

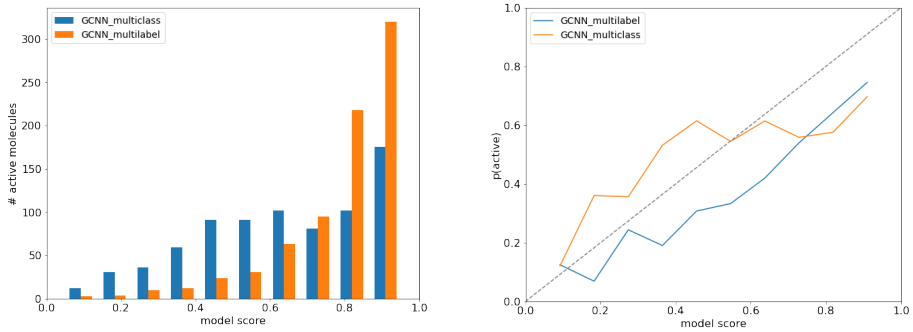
2.2 Wet-lab results

Improving hit-finding performance The training data follows the same preprocessing step with disynthon aggregation (Methods section of [1]), which is mainly to denoise the raw DNA-sequencing count data. During prospective experimental testing, each model type proposes a list of $\sim$ 200 molecules from the same commercially purchasable library (Mcule instock library [7] was used in this study) and the percentage of hits is measured in wet lab. We limit human intervention in the molecule list proposal by automating diversity selection and structural filtering into a streamlined pipeline, so that the hit-rate difference should be mostly explained by model type difference. Simplified from a two-step prospective testing approach [1], we use single-point inhibition assays evaluated at the molecule concentration of 10uM in this study.



(a) (b)

Figure 3: Prospective testing shows multilabel model outperformed multiclass model at all three inhibition cut-offs for (a) c-KIT target and (b) ERa target. The legend "paper" stands for the hit-rates reported in [1]. Note that for the multilabel model on ERa target we are still waiting to receive wet-lab test results of remaining 50% of the proposed compounds. The corresponding hit-rate is estimated from available wet-lab results so far.



(a) (b)

Figure 4: (a) Predictive probability distribution of hit molecules of an internal ERa inhibitor dataset using GCNN multiclass model and GCNN multilabel model. (b) is a reliability diagram for these two models

For both of the c-KIT and ERa protein targets, GCNN multilabel model outperformed GCNN multiclass model at three different inhibition cut-off percentages (see Figure 3). Specifically for c-KIT target, at 30% inhibition (enzyme remaining activity = 70%), multilabel model achieved 26.2% hit rate whereas multiclass model had 20.9%; at 30% inhibition, multilabel model achieved 20.1% hit rate whereas multiclass model had 17.4%. The hit rates achieved by multiclass model in [1] are also included as additional reference; the difference between multiclass model in this study and multiclass model in [1] come from recent improvements in model hyper parameters, difference in commercial libraries to select molecule list from, difference in diversity and filtering strategies.

Pushing hits to high score region Why does the GCNN multilabel model achieve a better hit-rate than the multiclass model? We notice that, via an internal ERa inhibitor dataset (test1) as a demonstrating example, GCNN multiclass model has placed many active molecules in the middle range of the score spectrum. This could be caused by unrealistic assumptions made in creating the mutually exclusive labels. On the other hand, GCNN multilabel model tends to score the active molecules (true inhibitors) more towards the high end compared with the GCNN multiclass model (see Figure 4a). Figure 4b suggests GCNN multilabel model has an improved calibration especially in the middle score region. This helps enrich the actives in the high score region.

3 Selecting key components

There are several key components that are essential for multilabel neural architecture to work properly. Below we disclose the our detailed choices in two important components: data sampling, label reduction scheme

Data sampling strategy We design the data sampling as follows. Each selection experiment label has a wide range of enrichment scores. We categorize each label into **high** enrichment and **low** enrichment types and feed the model during training with equal chance of high vs low enrichment data types. Furthermore, if we have 3 labels (i.e., 3 selection experiments), then the number of all combinatorial types is 8. We make sure each training data batch has balanced example counts between all these types.

Label reduction scheme Coming out of the GCNN multilabel model are multiple prediction scores (one per selection experiment), so how to effectively combine them into a single prediction (i.e., reduced label) and use it for checkpoint selection as well as downstream tasks (e.g., propose molecule list to find hits) are crucial for a successful multilabel model. Through *in-silico* evaluation, the final label reduction scheme used in this study is:

$$\text{Reduced-label} = \text{Target-Enr-label} - \text{NTC-Enr-label} - \text{Competitor-Enr-label} \quad (1)$$

4 Conclusion

In this study, we have presented a simple multilabel architecture for DEL data modeling, which allows models to learn more naturally with the DEL data at hand than the current state-of-the-art approach. An equally important aspect is that it helps eliminate the dependency on human-crafted column scheme. This simple multilabel architecture not only shows improvements in retrospective test datasets, but also, through web-lab prospective testing, brings real improvements in hit-finding use-case. Although the fundamental reason behind the effectiveness of multilabel models is yet to uncover, we observe that its improved calibration tends to score those active compounds towards higher region than the GCNN baseline model does.

References

[1] Kevin McCloskey, Eric A Sigel, Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A Clark, John W Cuozzo, Marie-Aude Guié, John P Guilinger, Christelle Huguet, Christopher D Hupp, Anthony D Keefe, Christopher J Mulhern, Ying Zhang, and Patrick Riley. Machine learning on DNA-Encoded libraries: A new paradigm for hit finding. *J. Med. Chem.*, June 2020.

[2] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.*, 30(8):595–608, August 2016.

[3] Matthew A Clark, Raksha A Acharya, Christopher C Arico-Muendel, Svetlana L Belyanskaya, Dennis R Benjamin, Neil R Carlson, Paolo A Centrella, Cynthia H Chiu, Steffen P Creaser, John W Cuozzo, Christopher P Davie, Yun Ding, G Joseph Franklin, Kurt D Franzen, Malcolm L Gefter, Steven P Hale, Nils J V Hansen, David I Israel, Jinwei Jiang, Malcolm J Kavarana, Michael S Kelley, Christopher S Kollmann, Fan Li, Kenneth Lind, Sibongile Mataruse, Patricia F Medeiros, Jeffrey A Messer, Paul Myers, Heather O'Keefe, Matthew C Oliff, Cecil E Rise, Alexander L Satz, Steven R Skinner, Jennifer L Svendsen, Lujia Tang, Kurt van Vloten, Richard W Wagner, Gang Yao, Baoguang Zhao, and Barry A Morgan. Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nat. Chem. Biol.*, 5(9):647–654, September 2009.

[4] Laura Guasch, Michael Reutlinger, Daniel Stoffler, and Moreno Wichert. Augmenting chemical space with DNA-encoded library technology and machine learning. *Chimia*, 75(1):105–107, February 2021.

[5] Péter Kómár and Marko Kalinić. Denoising DNA encoded library screens with sparse learning. *ACS Comb. Sci.*, 22(8):410–421, August 2020.

[6] Steven Kearnes. Pursuing a prospective perspective. *Trends in Chemistry*, November 2020.

[7] Robert Kiss, Mark Sandor, and Ferenc A Szalai. http://mcule.com: a public web service for drug discovery. *J. Cheminform.*, 4(1):1–1, May 2012.