Regression modeling on DNA encoded libraries

Ralph Ma[1]
insitro
279 E Grand Ave.,
South San Francisco, CA, 94080
ralphma@insitro.com

Gabriel H. S. Dreiman[1]
insitro
279 E Grand Ave.,
South San Francisco, CA, 94080
gdreiman@insitro.com

Fiorella Ruggiu[1]
insitro
279 E Grand Ave.,
South San Francisco, CA, 94080
fioruggiu@insitro.com

Adam J. Riesselman [2]
Hippo Harvest
adam@hippoharvest.com

Bowen Liu
insitro
279 E Grand Ave.,
South San Francisco, CA, 94080
bowen@insitro.com

Keith James
insitro
279 E Grand Ave.,
South San Francisco, CA, 94080
keith@insitro.com

Daphne Koller[3]
insitro
279 E Grand Ave.,
South San Francisco, CA, 94080
daphne@insitro.com

Mohammad M. Sultan[3]
insitro
279 E Grand Ave.,
South San Francisco, CA, 94080
msultan@insitro.com

Abstract

DNA encoded libraries (DELs) are pooled, combinatorial compound collections where each member is tagged with its own unique DNA barcode. DELs are used in drug discovery for early hit finding against protein targets. Recently, several groups have proposed building machine learning models with quantities derived from DEL datasets. However, DEL datasets have a low signal-to-noise ratio which makes modeling them challenging. To that end, we propose a novel graph neural network (GNN) based regression model that directly predicts enrichment scores from raw sequencing counts while accounting for multiple sources of technical variation and intrinsic assay noise. We show that our GNN regression model quantitatively outperforms standard classification approaches and can be used to find diverse sets of molecules in external virtual libraries.

[1]The three authors contributed equally to this paper.
[2]Previously employed by Insitro, Inc., where the research supporting this publication was conducted.
[3]Corresponding author

NeurIPS 2021 AI for Science Workshop.

Introduction

Small molecule drug discovery begins with the identification of putative chemical matter that binds to protein targets of interest. This can be be achieved with experimental techniques such as high-throughput screening or in silico methodologies such as docking and generative modeling. DNA encoded library (DEL)[11, 6, 2, 5, 4, 7] screening is a high throughput experimental technique used to identify diverse sets of chemical matter against targets of interest.

DELs are DNA barcode labeled pooled compound collections (Figure 1.A) that are incubated with an immobilized protein target in a process referred to as panning. The mixture is then washed to remove non-binders, and the remaining bound compounds are eluted, amplified, and sequenced to identify putative binders. DELs provide a quantitative readout for hundreds of millions of compounds. This readout can contain substantial experimental noise and biases caused by sources such as DEL members binding the protein immobilization media or differences in starting population (load). This noise is often controlled for by computing an enrichment metric[6, 11, 7, 5] that compares the compound population to its starting population and control experiments. Often, multiple replicates and off-target counter pans are used to improve the signal-to-noise ratio.

Recently, several groups [14, 12] have demonstrated powerful applications of ML to DELs. McCloskey et al. [14] trained classification models on labels generated with aggregated DEL counts and used the models to perform inference on large virtual libraries (VLs). This inference yielded diverse hits for several protein targets. However, their approach required determining a label boundary, which has no standard process and may bias the dataset. Moreover, classification models struggle to differentiate between weak and strong binders. This becomes important when prioritizing compounds for testing due to the multi-billion scale of VLs. Lim et al. [12] improved on the classification approach by directly modeling an enrichment metric (the ratio between counts from a target and an off-target pan). Their model was trained using a custom negative-log-likelihood loss function derived from a Poisson ratio test. However, it is not obvious that their method can be expanded to multiple off-target experiments or extended to include additional covariates.

By contrast, we introduce a general framework for quantitative modeling of DELs that incorporates multiple sources of experimental noise. We show that GNNs trained using our framework and regularized negative binomial loss outperform classification models in replicating experimental compound rankings generated with proxy measurements of binding affinity on a held out DEL. We end by examining key metrics from a virtual screen of the Enamine REAL and Wuxi VLs.

Methods

We obtained multiple proprietary DEL panning datasets screened against a challenging protein target, Interleukin 17A (IL-17A), a relevant target to autoimmune diseases. These datasets include several control and off-target pans. Here, we present results for a diversity screening library of 100M compounds (Lib1) that we used for training and a separate expansion library of 2.5M compounds used for validation (Lib2). The library synthesis, the panning experiments and the bioinformatics decoding and counting pipeline were performed at DICE Therapeutics. To start the synthesis, DNA barcodes are assembled. These barcodes have variable regions, codons, that encode the building blocks added to the synthesis and invariable regions that separate the codons. A unique molecular identifier(UMI) region is added to each DNA strand in the starting pool of barcodes. The UMI helps to de-bias the PCR amplification process, in which certain DNA barcodes may be more amplified than others. For libraries without UMIs, PCR bias can be a source of error. During library synthesis, chemical building blocks are added sequentially to each well and are encoded by their corresponding codon. All wells are then pooled back together before the the next building block is installed. Repeating this process creates a combinatorial chemical library with DNA barcodes corresponding to the chemical blocks added. However, chemical reactions have different yields and so the compounds attached to the same DNA encoding consists of a mixture of different products. Truncated product controls by omitting building blocks are often incorporated into the design of the library to ensure that a possible truncation is not what is binding. Furthermore, it is possible to have side products and racemization in the reactions. Combined, these factors confounds the count readout of each compound, making the readout a noisy signal for binding affinity. Additionally, some reactions and conditions may be DNA damaging, thereby reducing the DNA that can be decoded at the end of the panning procedure. To control for the overall imbalance in starting barcodes, the DNA barcodes are sequenced before the panning experiment (load counts). Finally, the library is incubated with the protein of interest.

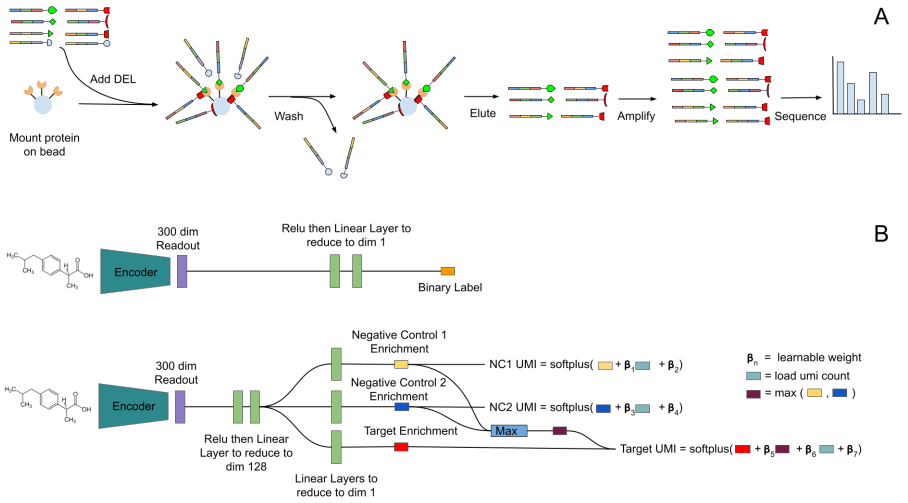

Figure 1: A. Diagram of DEL panning experiment. The sequencing counts can be attributed to both target specific binding (green) and non-target binding modes (red). B. Model schematics. Both classification and regression use a GIN-E encoder. The classification network maps the encoder output to a single class prediction. The regression network has multiple heads, each predicting an enrichment value from a reduced embedding of the encoder output. A linear sum of the enrichments and covariates is used to predict observed counts.

During this process, the DNA barcodes may stick to the protein (not a major risk for IL-17A) and the media of the experiment, including the beads and the protein's immobilization tags. This can be controlled by negative control pans which consist of running the panning experiment with the same experimental media, but without the protein or with an alternative control protein. Finally, problems with promiscuous compounds, compounds that may bind to multiple proteins, can be mediated by analyzing panning experiments for multiple proteins. Within our datasets, we had access to several such negative control panning data.

Classification Model

To provide a competitive baseline, we built and optimized a classification model using the same GNN architecture as the regression model GIN-E (Graph Isomorphism network with virtual node [15]). We assigned binary labels for binders (positives) and non-binders (negatives) using a two-step thresholding process. First, we discarded compounds with on-target unique molecular identifier (UMI) counts below a noise threshold. Second, we normalized compound UMIs in each pan by the sum of all UMIs in the pan to yield molecular frequencies (MFs). We then calculated the ratio between the on-target and max control or off-target MF. If a compound's MF ratio exceeded a positive cutoff or fell below a negative cutoff, we assigned it a positive or negative label, respectively. Compounds with ratios falling between the cutoffs were discarded. This yielded ~74K positives and ~5.6M negatives. We experimented with combinations of sampling schemes and losses to address the class imbalance, and found that Focal Loss [13] without balanced sampling performed best. Additionally, we regularized the model with dropout in the layers after graph readout and with input augmentations.

Regression Model

We used negative binomial regression to model the UMI from each panning experiment. This approach for denoising counts has previously been shown by [8] who used a regularized negative binomial regression model to de-noise UMI readouts from single cell RNA sequencing. They defined enrichment as the Pearson residual after regressing the count on sequencing depth. Similarly, we modeled the enrichment for each compound as the residual after accounting for various covariates such as binding to beads. As a generalization of Poisson regression, negative binomial regression incorporates a dispersion parameter α in addition to a mean variable μ. For one target pan and two no-

target control pans, $C_{i,target} \sim NB(\mu_{i,target}, \alpha_{target})$, $C_{i,control_1} \sim NB(\mu_{i,control_1}, \alpha_{control_1})$, and $C_{i,control_2} \sim NB(\mu_{i,control_2}, \alpha_{control_2})$ represent the UMI counts of ith compound in the respective panning experiments. We modeled μ_i as the combination of enrichment from binding to the target ($R_{i,target}$), enrichment from binding to the non-target media ($R_{i,control_1}$, $R_{i,control_2}$), and observed count of the compound in the original starting population load (S_i).

$$\mu_{i,control_1} = \sigma(R_{i,control_1} + \beta_1 S_i + \beta_2) \tag{1}$$

$$\mu_{i,control_2} = \sigma(R_{i,control_2} + \beta_3 S_i + \beta_4) \tag{2}$$

$$\mu_{i,target} = \sigma(R_{i,target} + \beta_5 max(R_{i,control_1}, R_{i,control_2}) + \beta_6 S_i + \beta_7) \tag{3}$$

β_i are learned from the data and σ represents the softplus function, which we found to be more stable during training than the typical exponential function. The dispersion parameter, α, of the negative binomial is a single scalar, learned for each experiment. We related $R_{i,target}$ and $R_{i,control}$ to each compound's structure by deriving their values with a GNN operating on the compound's molecular graph. A shared encoding network generates a 128 dimensional embedding vector from atom and bond features. This embedding vector is then transformed into $R_{i,target}$, $R_{i,control_1}$, and $R_{i,control_2}$ by separate feed forward networks (Fig 1.B). For our experiments, we used a GIN-E network with virtual node [15, 9, 10] for the initial encoding and two layers in each of our feed forward networks. During training, we summed the negative log likelihood of the observed counts for the target and control pans. Furthermore, we L2 regularized the enrichment values, which empirically prevented over-fitting. For a single example with count c_i for each panning experiment , the loss can be written:

$$P(c_i|\mu_i, \alpha) = \frac{\Gamma(c_i + \alpha^{-1})}{\Gamma(c_i + 1)\Gamma(\alpha^{-1})} \left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{c_i} \tag{4}$$

$$L_{i,target} = -\log P(c_{i,target}|\mu_{i,target}, \alpha_{target}) \tag{5}$$

$$L_{i,control_1} = -\log P(c_{i,control_1}|\mu_{i,control_1}, \alpha_{control_1}) \tag{6}$$

$$L_{i,control_2} = -\log P(c_{i,control_2}|\mu_{i,control_2}, \alpha_{control_2}) \tag{7}$$

$$L_i = L_{i,target} + L_{i,control_1} + L_{i,control_2} + \gamma R^2_{i,target} + \gamma R^2_{i,control_1} + \gamma R^2_{i,control_2} \tag{8}$$

where $\Gamma(x)$ is the gamma function and γ is the L2 regularization rate. This negative binomial regression can be further extended with other covariates such as enrichment in other negative control pans, other target pans, compound synthesis yield, and reaction type. For our experiments, we used 13 negative control pans. During validation and inference for virtual screening, we used our de-noised enrichment value $R_{i,target}$ to rank compounds.

Cross Library Validation

After training on Lib1, we validated our models on Lib2 which had proxy binding affinity measurements. Binding affinity of a compound to a target can be measured by the equilibrium disassociation constant Kd and corresponding negative log value pKd. Lib2 was used in a set of target titration panning experiments[3] where corresponding curves are fitted to produce a predicted titration-based pKds (t-pKds). A small portion of these t-pKds were validated with off-DNA pKd measurements (R^2 = 0.84). We measured model performance by calculating the Spearman correlation coefficient between model predictions and the t-pKds. This metric aligned with our intended use of the models to rank VLs for candidate selection.

The R_{target} predicted by our regression model had a 0.41 (95% CI [0.40, 0.43]) Spearman correlation with t-pKds (Fig 2.A). This exceeds both a Random Forest classification baseline (0.28) and our GNN classification model (0.35 (95% CI [.34,.37])) (Fig 2.B). Furthermore, both the GNN regression and GNN classification models trained from Lib1 showed better correlation with t-pKds from Lib2 than UMI counts from a single pan of Lib2. This illustrates both the high noise in the raw UMI output from a single panning experiment and our models' ability to generalize. Finally, we compared the two GNN models' retrieval rates for strong binders in their top prediction results. The regression model had more binders in its top prediction results than the classification model (Fig 2.C).

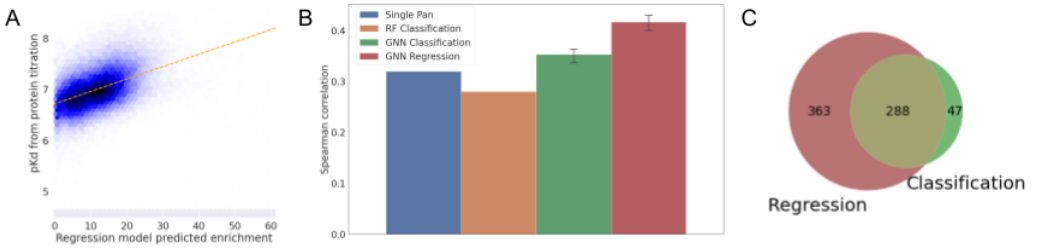

Figure 2: A. Bivariate histogram showing correlation between predicted enrichment (R_{target}) from GNN regression model and t-pKds derived from protein titration. Regression line is plotted in orange. B. Spearman correlation between model predictions and t-pKds. For RF Classification and GNN Classification, predicted probability of the binder class is used for ranking. "Single Pan" represents the UMI counts from a single panning experiment done with Lib2 at the same target concentration as experiments with Lib1. Error bars on GNN Regression and Classification represents the 95% confidence interval as estimated by three model replicates. C. Venn diagram showing number of compounds with t-pKds >= 8 (n=1327) retrieved in the top 10K of the GNN Regression versus the GNN Classification model.

Virtual Screening

We used our regression model to perform a virtual screen of 3.7 billion compounds in Enamine and WuXi's VLs. We then thresholded the regression model's output to select a limited number of top scoring compounds for analysis. We established the threshold using a quantitative analysis of the enrichment distribution in both the validation and inference sets (Fig 3.A). We clustered this top set using Taylor-Butina clustering[1] with a similarity cut-off of 0.6. Structural similarity was calculated via Jaccard similarity of Morgan fingerprints. This clustering yielded 26 clusters, and we selected top scoring compounds from each cluster for testing. These selected compounds are structurally diverse (Fig 3.C).

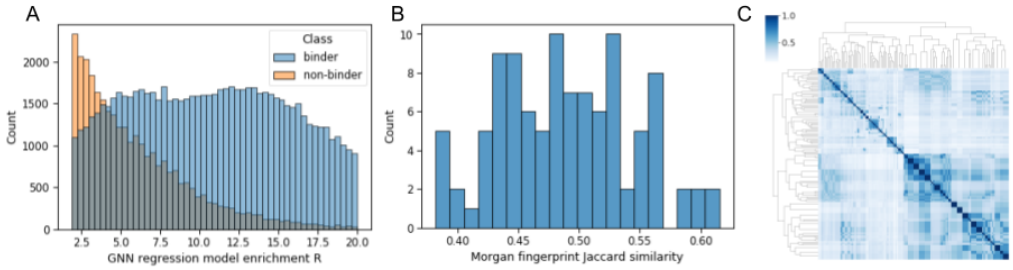

Figure 3: Regression models pick diverse compounds from the VLs. A. Histogram of predicted enrichment for the overlapping region between binders and non-binders on the validation set (Lib2). B. Histogram of Jaccard similarities for inference compounds with predicted enrichment >10 to their closest neighbor in the training set (Lib1). C. Heatmap of pairwise Jaccard similarities between compounds with predicted enrichment >10.

Conclusion

DEL experiments yield datasets with low signal-to-noise ratio. In our work, we show a novel regression technique for modeling DEL sequencing counts that accounts for various sources of variation, such as media binding and differences in initial load. Our model's predicted enrichment values have better correlation with proxy binding affinities than those of baseline classification models or experimental values from a single panning experiment. Finally, we demonstrate that our model retrieves diverse compounds during virtual screening.

Acknowledgments

The authors would like to acknowledge the team at DICE therapeutics for graciously providing the DEL data used in this paper. We would also like to thank multiple insitro colleagues for their help in directly and indirectly making this manuscript possible.

References

[1] D. Butina. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 07 1999.

[2] M. A. Clark, R. A. Acharya, C. C. Arico-Muendel, S. L. Belyanskaya, D. R. Benjamin, N. R. Carlson, P. A. Centrella, C. H. Chiu, S. P. Creaser, J. W. Cuozzo, C. P. Davie, Y. Ding, G. J. Franklin, K. D. Franzen, M. L. Gefter, S. P. Hale, N. J. V. Hansen, D. I. Israel, J. Jiang, M. J. Kavarana, M. S. Kelley, C. S. Kollmann, F. Li, K. Lind, S. Mataruse, P. F. Medeiros, J. A. Messer, P. Myers, H. O'Keefe, M. C. Oliff, C. E. Rise, A. L. Satz, S. R. Skinner, J. L. Svendsen, L. Tang, K. van Vloten, R. W. Wagner, G. Yao, B. Zhao, and B. A. Morgan. Design, synthesis and selection of dna-encoded small-molecule libraries. *Nature Chemical Biology*, 5(9):647–654, 2009.

[3] J. Cuozzo, P. Centrella, D. Gikunju, S. Habeshian, C. Hupp, A. Keefe, E. Sigel, H. Soutter, H. Thomson, Y. Zhang, and M. Clark. Discovery of a potent btk inhibitor with a novel binding mode using parallel selections with a dna-encoded chemical library. *Chembiochem : a European journal of chemical biology*, 18:864–871, 01 2017. doi: 10.1002/cbic.201600573.

[4] W. Decurtins, M. Wichert, R. M. Franzini, F. Buller, M. A. Stravs, Y. Zhang, D. Neri, and J. Scheuermann. Automated screening for small organic ligands using dna-encoded chemical libraries. *Nature Protocols*, 11(4):764–780, 2016.

[5] J. C. Faver, K. Riehle, D. R. Lancia, J. B. J. Milbank, C. S. Kollmann, N. Simmons, Z. Yu, and M. M. Matzuk. Quantitative comparison of enrichment from dna-encoded chemical library selections. *ACS Combinatorial Science*, 21(2):75–82, 02 2019.

[6] C. J. Gerry, M. J. Wawer, P. A. Clemons, and S. L. Schreiber. Dna barcoding a complete matrix of stereoisomeric small molecules. *Journal of the American Chemical Society*, 141(26): 10225–10235, 07 2019.

[7] A. Gironda-Martínez, E. J. Donckele, F. Samain, and D. Neri. Dna-encoded chemical libraries: A comprehensive review with successful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279, 08 2021.

[8] C. Hafemeister and R. Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, 2019.

[9] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec. Strategies for pre-training graph neural networks, 2020.

[10] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open graph benchmark: Datasets for machine learning on graphs, 2021.

[11] L. Kuai, T. O'Keeffe, and C. Arico-Muendel. Randomness in dna encoded library selection data can be modeled for more reliable enrichment calculation. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 23(5):405–416, 2021/09/07 2018.

[12] K. S. Lim, A. G. Reidenbach, B. K. Hua, J. W. Mason, C. J. Gerry, P. A. Clemons, and C. W. Coley. Machine learning on dna-encoded library count data using an uncertainty-aware probabilistic loss function, 2021.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection, 2018.

[14] K. McCloskey, E. A. Sigel, S. Kearnes, L. Xue, X. Tian, D. Moccia, D. Gikunju, S. Bazzaz, B. Chan, M. A. Clark, J. W. Cuozzo, M.-A. Guié, J. P. Guilinger, C. Huguet, C. D. Hupp, A. D. Keefe, C. J. Mulhern, Y. Zhang, and P. Riley. Machine learning on dna-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866, 08 2020. doi: 10.1021/acs.jmedchem.0c00452. URL https://doi.org/10.1021/acs.jmedchem.0c00452.

[15] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks?, 2019.