Hit Expansion Driven By Machine Learning

Jin Xu
Google Research
jix@google.com

Steven Kearnes *
Relay Therapeutics
skearnes@relaytx.com

Jianwen A. Feng †
Google Research
jw.a.feng@gmail.com

Abstract

Recent work [1] utilized experimental data from DNA-encoded library (DEL) selections to train graph convolutional neural networks (GCNNs) [2] for identifying hit compounds for protein targets and their prospective test results demonstrated excellent hit rates for three diverse proteins. Building on this work, we propose two novel approaches to leverage DEL GCNN model predictions and embeddings to automate hit expansion, a critical step in real-world drug discovery that guides the optimization of initial hit compounds toward clinical candidates. We prospectively tested the proposed approaches on a protein target (sEH) and our methods identified more small molecules with higher potency compared to traditional molecular fingerprint similarity searches. Specifically, we discovered 34 molecules with higher potency than a sEH clinical trial candidate using our approaches. All sEH assay results are publicly available at `https://www.tdcommons.org/dpubs_series/6300`. Furthermore, applying the automated hit expansion approach to WDR91, a novel protein target that has no known binders, led to the discovery of two first-in-class covalent binders that were experimentally confirmed by co-crystal structures.

1 Introduction

Small molecule drug discovery involves identifying novel therapeutic compounds to treat diseases through successive optimization of initial hit compounds. The process typically begins with high-throughput screening to find hits that are active against a biological target. These initial hits then go through hit expansion, synthesizing and testing chemical analogs to expand active chemical series. The most promising compounds advance to hit-to-lead and lead optimization phases to enhance potency, selectivity and drug-like properties. Traditionally, medicinal chemists manually carry out successive rounds of chemical optimization, relying heavily on their expertise and knowledge to iteratively synthesize and modify compounds. As such, there is increasing interest in using computational approaches to automate portions of these chemical optimization cycles to progress compound refinement efficiently. This work presents automated hit expansion approaches utilizing a predictive model trained on DNA-encoded library (DEL) selection data.

DEL screening [3, 4] has emerged as an efficient and cost-effective alternative to high-throughput screening (HTS) for hit finding. A recent work [1] introduced a new method for identifying hit compounds using DEL screening without the need for expensive custom off-DNA chemical synthesis. They trained graph convolutional neural networks (GCNNs) for affinity prediction using the affinity-mediated DEL selection output data. These GCNN models were then utilized to perform virtual screening on commercially available compound catalogs as illustrated in Figure 1(b) of [1]. This approach achieved excellent hit rates for three diverse protein targets.

*Work done in Google Research
†Work done in Google Research

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

However, initial hit identification is only the first step. In real-world drug discovery, hits are followed by hit expansion, hit-to-lead, and lead optimization phases that chemically modify the hits to optimize them into clinical candidate compounds. These phases differ from hit-finding by: (1) a focus on local multi-objective optimization of chemical space (rather than the discovery of diverse starting points), and (2) generation of small amounts of new experimental data (10s to 100s of compounds) to guide multiple rounds of optimization.

Extending the work of [1], this paper explores an interesting question: can the predictive GCNN model trained with DEL screening data work in hit expansion? In other words, can the DEL model, originally trained for hit identification, be used to guide local search and prioritize analogs of confirmed hit compounds? We propose two approaches to utilize the GCNN model: first, directly using GCNN prediction scores; and second, training a secondary model based on GCNN embeddings using $O(100)$ new experimental data points. To evaluate these approaches, we prospectively assessed predictions by purchasing molecules and experimentally testing them against soluble epoxide hydrolase (sEH). For comparison, we included a baseline method using molecular fingerprint similarity. After obtaining strong performance of our approaches with sEH, a target with known small molecule inhibitors, we successfully applied them to a novel protein target WDR91 [5], a challenging under-explored target without known small molecule binders.

2 Methodology

Given a GCNN model trained on DEL screening data for a protein target of interest as shown in Figure 1(b) of [1], this section describes the proposed approaches for machine learning guided hit expansion, outlined in the following steps.

2.1 Analog search in Enamine REAL

Hit expansion was conducted by searching the Enamine REAL virtual library, which contains 1.9 billion commercially available synthesize-on-demand compounds. Enamine REAL molecules do not require expensive and time consuming bespoke synthesis, making them cost-effective for hit-expansion.

For each starting hit compound, we performed similarity searches within the 1.9 billion Enamine REAL library to find analogs with extended-connectivity fingerprints with radius 3 (ECFP6) [6] Tanimoto similarity exceeding a specific threshold (0.4). This process resulted in a set of analogs defining a local chemical space that the subsequent hit-expansion methods could search.

2.2 GCNN Prediction-Guided Hit Expansion

This approach directly utilized the GCNN predictions to guide hit expansion. We inferred GCNN scores for all analog compounds identified through the similarity search method described in Section 2.1. The GCNN prediction scores were then used to rank the analogs. To balance exploration and exploitation, we also applied a directed sphere exclusion (DISE) algorithm [7] to select the top-scoring molecules from each cluster for purchasing. To evaluate this GCNN prediction-guided approach, we compared it against a baseline method of ranking the analogs based on ECFP6 Tanimoto similarity to the starting hit compounds.

2.3 Secondary DEL Model-Guided Hit Expansion

Purchasing and testing analog compounds proposed by the GCNN prediction-guided hit expansion in Section 2.2 generates hundreds of new data points with accurate and precise experimentally measured potency values. In this work, dose response potency values are reported as pIC_{50} which is the negative log of the IC_{50} (the molar concentration producing 50% inhibition). This is in contrast to the noisy, large-scale DEL selection data comprising millions of diverse molecules. Our goal is to leverage this locally generated, accurate potency data to train a secondary model to improve the potency prediction performance in searching chemical space around initial hit compounds.

Our proposed approach of training a secondary model, as illustrated in Figure 1, treats the last layer of the trained GCNN as a molecular representation. The GCNN model was originally trained in a classification context for hit-finding on the protein target of interest, and we hypothesized that these

learned embeddings may outperform traditional molecular fingerprints such as extended-connectivity fingerprints (ECFP) [6] for predicting potency on that same target. In our approach, the GCNN model was frozen and the GCNN embeddings were utilized as input features for a support vector machine (SVM) model. This SVM model was then trained to predict the pIC_{50} using hundreds of new experimental pIC_{50} data points of the analog compounds (acquired from the first round of hit expansion described in Section 2.2, including compounds selected by both the GCNN and the similarity search) as the ground-truth training data. In this way, we can take advantage of the representations learned by the GCNN on the hit-finding task and adapt them, through a secondary model, for the related task of potency prediction, using the new experimental pIC_{50} data of selected analogs.

After training, we replaced the GCNN predictions used in Section 2.2 with predictions from the secondary model to guide subsequent hit expansion, using a similar process as described in Section 2.2.

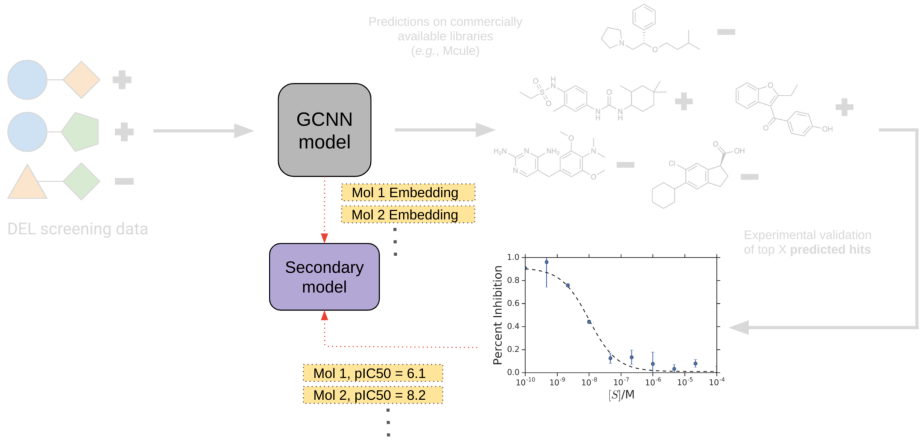

Figure 1: A secondary potency prediction model, trained on frozen GCNN molecular embeddings, is used to predict compound pIC_{50} values. This figure builds upon the hit finding workflow diagram in Figure 1(b) of [1].

3 Experiments

To compare the hit expansion approaches, we purchased predicted molecules and prospectively tested them against soluble epoxide hydrolase (sEH) in wet lab experiments. sEH has a known small molecule inhibitor, GSK2256294A [8], that has reached clinical trials and can be used as a benchmark compound in this hit expansion study.

3.1 Hit Finding for sEH

To prepare for hit expansion experiments, we first ran hit finding against sEH. We retrained the GCNN model on the sEH DEL selection data from [1], with some improvements such as migrating from CPUs to TPUs and ensembling replicated model runs to obtain robust predictions. Prioritized by the retrained GCNN predictions along with diversity and property filters, we ordered 200 diverse compounds from the Mcule instock catalog (9.35 million compounds), with 160 successfully delivered. Through dose response testing of these compounds, we discovered 128 active hits with $\text{pIC}_{50} \geq 6$, i.e., better than 1 μM IC_{50}.

Ultimately, we selected 12 hit compounds to serve as starting points for the following hit expansion study. These compounds were chosen based on top rankings across three metrics: potency (pIC_{50}), ligand-lipophilicity efficiency (LLE) defined as $\text{pIC}_{50} - \log \text{P}$, and ligand efficiency (LE) defined as $\text{pIC}_{50}/\text{HA}$ where HA stands for the number of heavy (non-hydrogen) atoms in the molecule. It is worth noting that, as a given compound could appear at the top of multiple ranked list from different metrics, we utilized a rotational picking approach when selecting the set of starting points. Specifically, we were picking one hit compound per each metric in a round robin fashion, until we reached the desired number of total starting point hits.

3.2 Hit Expansion Round 1 for sEH: GCNN Prediction-Guided Hit Expansion

In the first round of hit expansion, we searched for analog compounds in the Enamine REAL library of 1.9 billion compounds, requiring an ECFP6 Tanimoto similarity greater than 0.4 to the 12 initial hits. This resulted in 22470 analog compounds being retrieved. From these analogs of each hit, we selected the highest GCNN scoring compound in each of the top 10 DISE clusters. As a baseline, we also selected the top 10 compounds with the highest ECFP6 similarity to each of the 12 initial hits. This process resulted in 120 analog compounds for the baseline set and 120 analog compounds for our proposed ordering. In total, 219 compounds were ultimately synthesized for experimental testing.

Each compound underwent dose response testing twice. The arithmetic mean of the two pIC_{50} values (pIC_{50}_mean) was used as the final testing score for each compound. Due to the 10 μM upper limit on testing, the pIC_{50} values for any inactive compounds set to 5. Figure 2 shows the prospective testing results as a complementary cumulative distribution function (CCDF) plot of the pIC_{50}_mean values. The CCDF plot shows that our proposed approach substantially outperformed the baseline. Specifically, compared to the baseline, our approach yielded 22% more compounds with IC_{50} < 100 nM, 12% more with IC_{50} < 10 nM, and 3% more with IC_{50} < 1 nM.

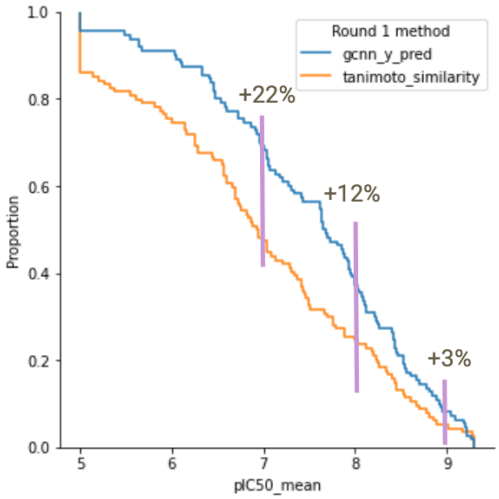

Figure 2: Round 1 hit expansion comparison, where gcnn_y_pred is GCNN prediction guided hit expansion and tanimoto_similarity is the baseline of using molecular fingerprint similarity.

Additionally, we compared the potency improvement of the analog compounds relative to their respective starting hit compounds. Our proposed approach also outperformed the baseline in terms of the number of analog compounds with potency improvement, as shown in Table 1.

Table 1: The number of analog compounds which has higher potency than their corresponding starting hit compounds.

	gcnn_y_pred	tanimoto_similarity
pIC_{50}(analog) − pIC_{50}(starting hit) > 0	**23**	13
pIC_{50}(analog) − pIC_{50}(starting hit) > 1	**6**	5

3.3 Hit Expansion Round 2 for sEH: Secondary DEL Model-Guided Hit Expansion

After the first round of hit expansion, we had accumulated potency testing data for 379 compounds, consisting of 160 from hit finding and 219 from hit expansion. We utilized these 379 pIC_{50} data points as ground truth labels and the GCNN embeddings as input features to train an SVM model. This secondary SVM model was then used to guide subsequent hit expansion, following a similar process to that described previously for the first round as in Section 3.2.

For round 2 starting hit compounds, we selected 15 compounds using slightly different criteria than round 1. These compounds were chosen based on top rankings across three metrics: having improved pIC_{50} by at least 1 unit in the first round hit expansion, having the best potency (pIC_{50}), or having the best ligand-lipophilicity efficiency (LLE) defined as $pIC_{50} - \log P$. This resulted in 2 hit compounds from hit finding round and 13 analog compounds from round 1 hit expansion being chosen as starting compounds. In this round, we used the GCNN prediction based hit expansion approach instead of the lower performing fingerprint similarity method as the baseline. This process resulted in 150 analog compounds for the baseline set and 150 analog compounds for our proposed approach of training a secondary model. After removing duplicates, 188 compounds were ultimately synthesized for experimental testing.

The CCDF plot in Figure 3 showed our proposed approach of training a secondary model greatly outperformed the baseline. With our approach, most of the ordered analog compounds had a high potency to sEH with pIC_{50}_mean > 8, meaning an $IC_{50} < 10$ nM. Specifically, compared to the baseline, our approach yielded 27% more compounds with $IC_{50} < 10$ nM and 23% more with $IC_{50} < 1$ nM.

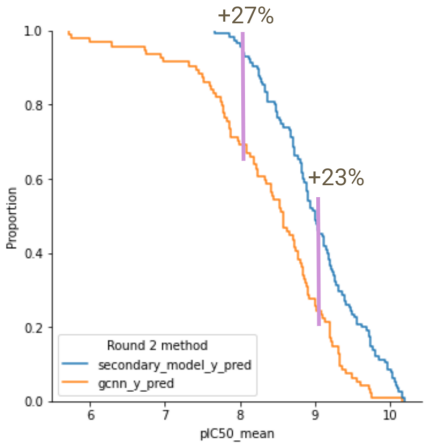

Figure 3: Round 2 hit expansion comparison, where secondary_model_y_pred is the proposed secondary DEL model guided hit expansion, and gcnn_y_pred is GCNN prediction guided hit expansion as the baseline.

Additionally, as shown in Table 2, the potency improvements demonstrated that our secondary model approach achieved substantially greater performance compared to the original GCNN model.

Table 2: The number of analog compounds which have higher potency than their corresponding starting hit compounds.

	secondary_model_y_pred	gcnn_y_pred
pIC_{50}(analog) $-$ pIC_{50}(starting hit) > 0	**57**	28
pIC_{50}(analog) $-$ pIC_{50}(starting hit) > 1	**6**	1

Furthermore, since we discovered so many high potency compounds in this round, we cherry-picked 62 of the top potent compounds and performed more fine-grained dose-response experiments on them. For this, we used a higher top concentration of 0.05 μM. From this, we discovered 34 analogs with higher potency than the sEH clinical candidate GSK2256294A. Notably, 25 of these came from the secondary model predictions approach (secondary_model_y_pred) while 9 came from the original GCNN predictions (gcnn_y_pred). This again demonstrated the improved performance obtained by training a secondary model on local analog data with GCNN embeddings versus the initial GCNN model.

Finally, to enable further research building on this work, we have made all sEH IC_{50} assay results publicly available at `https://www.tdcommons.org/dpubs_series/6300`.

3.4 Hit Expansion for WDR91: GCNN Prediction-Guided Hit Expansion

After obtaining excellent performance of our hit expansion approaches on sEH, which has known small molecule inhibitors, we applied this approach to the novel target WDR91, as reported in [5], which presents a greater challenge due to the absence of any known small molecule binders. This automated hit expansion led to the discovery of covalent analogs 18 and 19 as described in [5], which were confirmed to form a covalent adduct by experimental co-crystal structures. The discovery of these 2 active analogs provided valuable insights about structure-activity relationships (SAR) to guide medicinal chemistry efforts and accelerate the development of novel chemical tools to evaluate WDR91's therapeutic potential. This further demonstrated the effectiveness of the GCNN-guided approach for hit expansion on novel targets, in addition to those with known inhibitors such as sEH.

4 Conclusion

In this work, we proposed two approaches to leverage GCNN hit-finding models trained on DEL screening data to automate hit expansion from readily accessible and low-cost Enamine compound catalogs. Prospective testing showed that these techniques can discover more analog compounds with higher potency, compared to traditional fingerprint based similarity search. The first approach of GCNN prediction-guided hit expansion succeeded for both sEH, which has known inhibitors, and the challenging novel target WDR91, which previously had no known chemical probes. Training a secondary DEL model also yielded high performance on sEH, resulting in 34 analogs with higher potency than the clinical candidate GSK2256294A. However, it is worth noting that this approach required sufficient active compounds for model training. While promising for well-studied targets such as sEH, more development may be needed to enable effective training of a secondary model for novel proteins with smaller amounts of active compounds available. Overall, this study demonstrated DEL-trained GCNNs, with and without refinement on new data, can drive cost-efficient and effective hit expansion.

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] Sydney Brenner and Richard A. Lerner. Encoded combinatorial chemistry. *PNAS*, 89:5381–5383, June 1992.

[4] 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.

[5] Shabbir Ahmad, Jin Xu, Jianwen A Feng, Ashley Hutchinson, Hong Zeng, Pegah Ghiabi, Aiping Dong, Paolo A Centrella, Matthew A Clark, Marie-Aude Guié, John P Guilinger, Anthony D Keefe, Ying Zhang, Thomas Cerruti, John W. Cuozzo, Moritz von Rechenberg, Albina Bolotokova, Yanjun Li, Peter Loppnau, Alma Seitova, Yen-Yen Li, Vijayaratnam Santhakumar, Peter J. Brown, Suzanne Ackloo, and Levon Halabelian. Discovery of a first-in-class small molecule ligand for wdr91 using dna-encoded chemical library selection followed by machine learning. *https://doi.org/10.1101/2023.08.21.552681*, 2023.

[6] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J Chem Inf Model*, 50:742754, 2010.

[7] Alberto Gobbi and Man-Ling Lee. Dise: Directed sphere exclusion. *J Chem Inf Comput Sci*, 43 (1):317–323, 2003.

[8] Patricia L Podolin, Brian J Bolognese, Joseph F Foley, Edward Long 3rd, Brian Peck, Sandra Umbrecht, Xiaojun Zhang, Penny Zhu, Benjamin Schwartz, Wensheng Xie, Chad Quinn, Hongwei Qi, Sharon Sweitzer, Stephanie Chen, Marc Galop, Yun Ding, Svetlana L Belyanskaya, David I Israel, Barry A Morgan, David J Behm, Joseph P Marino Jr, Edit Kurali, Mary S Barnette, Ruth J Mayer, Catherine L Booth-Genthe, and James F Callahan. In vitro and in vivo characterization of a novel soluble epoxide hydrolase inhibitor. *Prostaglandins Other Lipid Mediat.*, 25-31:104–105, 2013.