Unveiling Hidden Treasures: Exploring Large Chemical Spaces with Machine Learning Models Trained on DNA-Encoded Libraries Selection Results

A. Kapeliukha¹, O. Strapak¹, O. Tarkhanova¹, S. Zozulya³, Alla Pohribna³, Yurii S. Moroz²

¹Chemspace LLC, 85 Winston Churchill Street, Kyiv 02094, Ukraine ²Enamine Ltd, 78 Winston Churchill Street, Kyiv 02094, Ukraine

³Bienta, 78 Winston Churchill Street, Kyiv 02094, Ukraine

DEL Screening Data

DEL screening results:

- 4 experiment selections
- 2 control selections
- 108.5k datapoints



Noise Reduction Full DEL compound A1 Fragment A1 substructures responsible for binding get higher average disynthons inherit copy count: 20 full molecule counts average: 5.25 counts average: 21 separate into 3 pair the fragments average across the dataset fragments together

Regression Model Performance

Model	Spearman Ki (full)	Spearman Ki (subset)
Molecular weight Benzenesulfonamide presence	-0.121 -0.199	0.074 -0.063
Top Vina docking score Top GNINA docking score	-0.068 -0.279 ± 0.044	0.119 -0.091 ± 0.061
RF trained on enrichment scores from Gerry et al.	-0.231 ± 0.007	-0.091 ± 0.012
GNN (Lim et al.)	-0.298 ± 0.005	-0.075 ± 0.011
DEL-Dock (Shmilovich et al.)	-0.328 ± 0.01	-0.186 ± 0.01
Our model	-0.309 ± 0.007	-0.214 ± 0.014

Full molecules - prediction on validation set

Spearman's r: 0.28 ± 0.01

Disynthons - prediction on validation set

Spearman's r: 0.59 ± 0.01

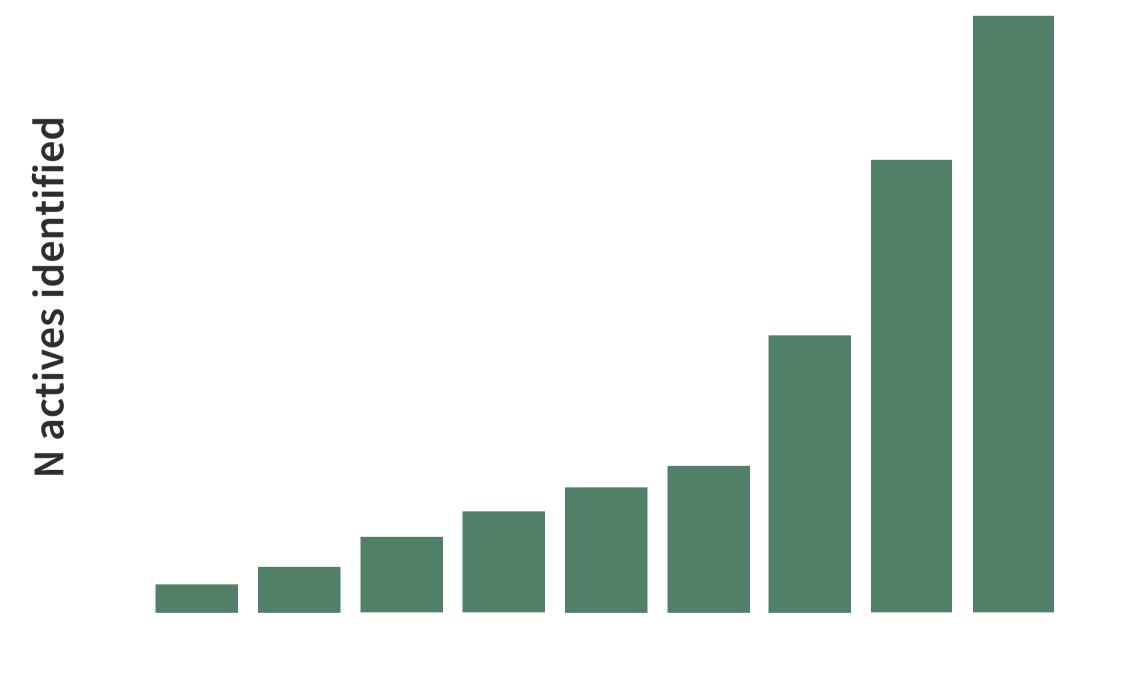
Predicted enrichment

Predicted enrichment

Exploration of Enamine REAL

Selecting active compounds from ChEMBL

Number of active compounds in top N ChEMBL ranked by predictions



Top N ranked ChEMBL compounds

9

33B+

Top compounds by **Enamine REAL** prediction Space

Sphere exclusion clustering

Visual inspection

- Regression model with best performance used for prediction
- Selected diverse compounds with the best predicted enrichment in the cluster
- Compounds are dissimilar from known actives
- Synthesis success rate 87% by Enamine

150 Selected for synthesis

130 Successfully synthesized

Results

- Disynthon aggregation helps to identify substructures important for binding and de-noise DEL datasets
- Regression models trained on DEL data can efficiently rank compounds from large compound databases.
- Exploration on Enamine REAL by ML model leads to identification of 33 novel hits with 25.4% hit rate.



a.kapeliukha@chem-space.com



