



BELKA

The big encoded library for chemical
assessment

NeurIPS 2024



BELKA Goal

Current datasets

- too small
- critical data bias issues
- fail to test binding prediction

Release large dataset to test if ML can **learn to generalize** binding predictions to new chemical space



BELKA in context

Predicting **P(pose | binds)** is much easier than predicting **P(binds)**

We made BELKA to help learn to predict **P(binds)**

Binding Pose Datasets

- **PDBBind**
 - ~20k protein-molecule pairs
- **POSEBUSTERS**
 - ~500 protein molecule pairs
 - Problems with splits, known data leaks
- **PLINDER**
 - ~450k protein-molecule pairs
 - Great Dataset, Much better splits, a bit hard to use



BELKA in context

Models latch onto dataset bias instead of learning to generalize

To limit bias, we wanted to create a large dataset from a single source/assay, without preselecting compounds.

Binding Datasets

- **PubChem**
 - Substantial publication and sampling bias limit usefulness of data. Requires great care to filter data
- **BindingDB**
 - ~1m data points
 - Better filters but still substantial problems with bias.



Splitting in chemistry

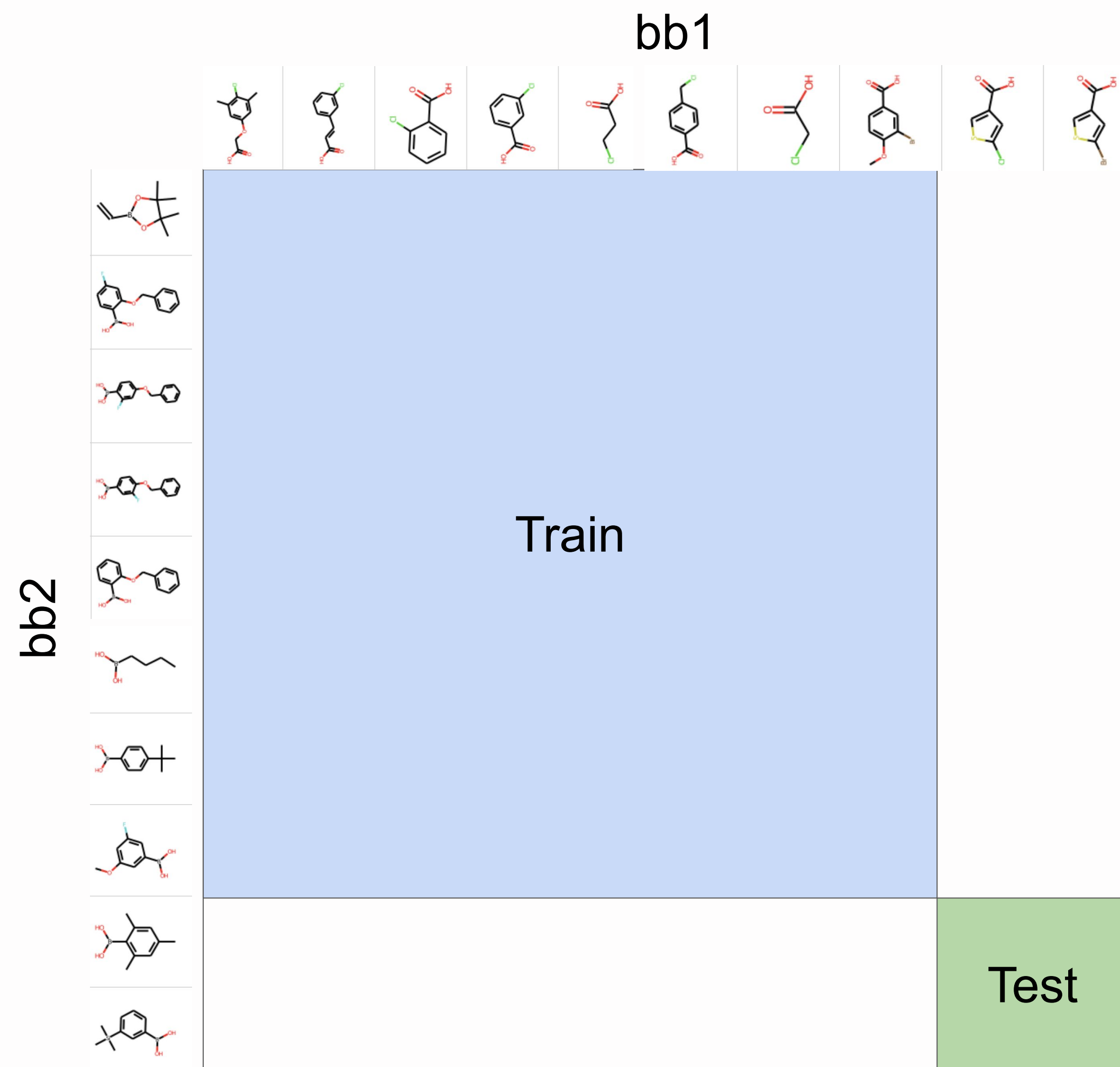
Models latch onto memorizing chemical motifs from training set instead of learning to generalize

Methods

- Random
- Scaffold
- Cluster based
- Building Block
- New Library



Building Block Split



Split so that no building blocks are shared between training set and test set.

Many molecules are lost with building block splits.



BELKA splits

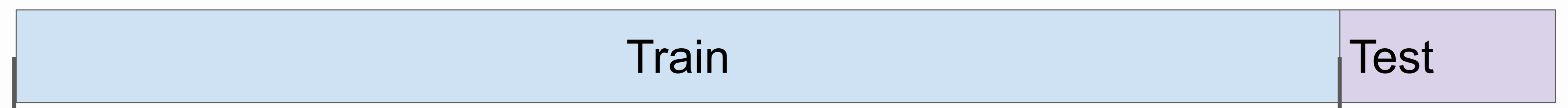
Library Split



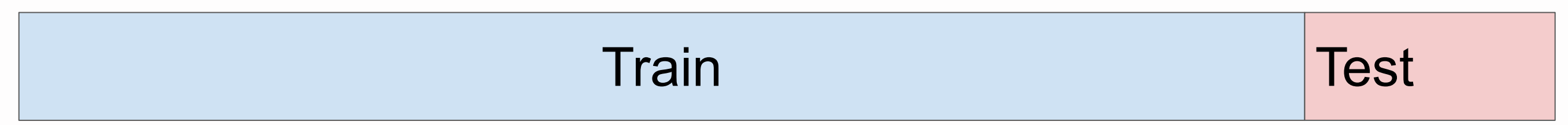
BB Split



Scaffold Split

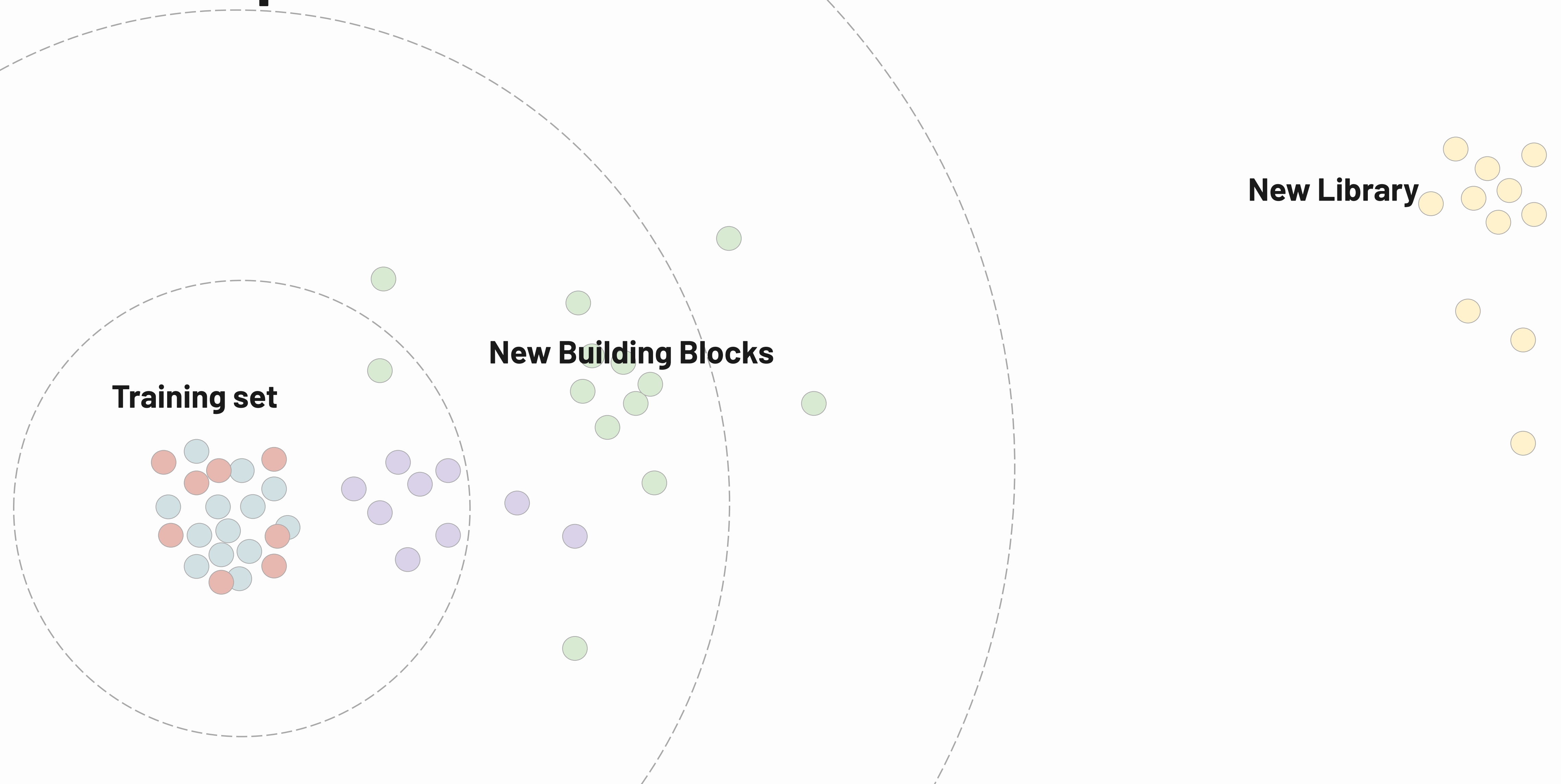


Random Split





BELKA split distributions





Metrics

Classification metrics

Accuracy

- Dataset is way too imbalanced

Average Precision

- Good method for testing model ability to rank

precision @ top 100

- closest to real life application
- perhaps too noisy for competition

We chose Average Precision for BELKA



Metrics - Lessons Learned

We chose Average Precision but made a mistake.

We calculated Average Precision over all groups together.



After realizing this we changed the competition metric to be calculated for each (split, protein) group separately and then averaged.



Kaggle results: participant's models mostly memorize and fail to generalize

KAGGLE CONTESTANTS

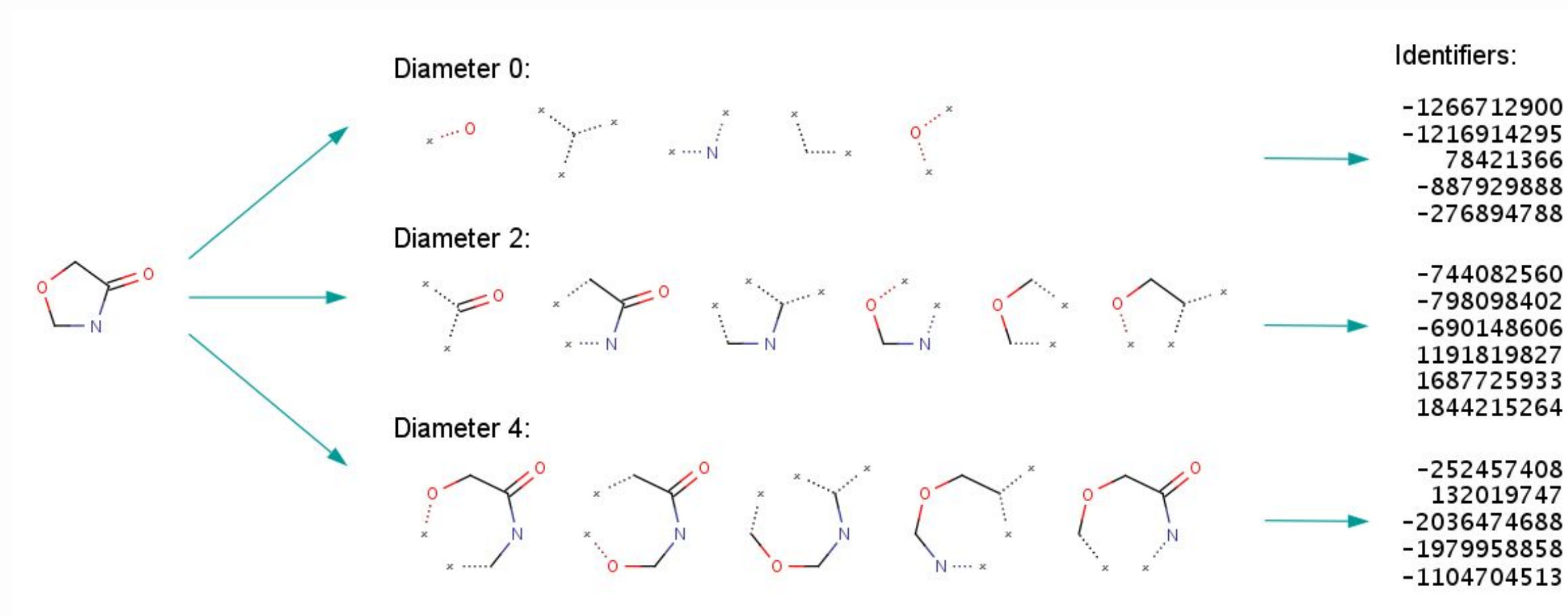
- Do well on motifs shared in the training set
- Do less well on non-shared motifs
- Don't generalize at all to new libraries

TECHNIQUES TRIED

- **ECFPs + tree based methods**
 - Good baseline, winning models did not outperform RF by very much
- **1DCNNs**
 - Unexpected wins - Won 3 of 5 prizes
 - High variance / Doesn't overfit as much
- **BERT or RoBERTa based language models**
 - Competition workhorse
 - Scales to full dataset
- **GNNs, GraphConvns or MPNNs**
 - Contestants had trouble with scaling and overfitting
- **Docking**
 - Not attempted much (due to cost)
 - reports that docking score did not correlate with binding

Winners were very careful on evaluating overfitting using custom data splits

Tree-based methods and ECFPs



Basically Hashing trick on subgraphs. Learns which chemical motifs correspond to binding.

Method can only memorize never generalize, but forms a very robust baseline



1DCNNs

Method really performed much better than expected.

Very good/popular tutorial released, which had high variance.

This method didn't seem to overfit as much as other methods.












1DCNNs

#	△	Team	Members		Score	Entries	Last	Solution
1	▲ 850	Victor Shlepov		●	0.30619	34	6mo	
2	▲ 1020	Smeet159		●	0.30259	2	6mo	
3	▲ 1023	Mohib		●	0.29028	9	6mo	
4	▲ 1004	Takako		●	0.28772	10	6mo	
5	▲ 37	mamba1-one-fold-lb0.432		●	0.28557	175	6mo	



1DCNNs

#	△	Team	Members		Score	Entries	Last	Solution
1	▲ 850	Victor Shlepov			0.30619	34	6mo	
2	▲ 1020	Notebook BELKA 1DCNN Starter with all data Version 1			0.30259	2	6mo	
3	▲ 1023	Notebook BELKA 1DCNN Starter with all data Version 4			0.29028	9	6mo	
4	▲ 1004	Notebook BELKA 1DCNN Starter with all data Version 2			0.28772	10	6mo	
5	▲ 37	mamba1-one-fold-lb0.432			0.28557	175	6mo	



BERT/Mamba

Models are pre-trained with a masking task.

Fast to train and provides good embeddings.



Graph Convs

Not many contestants use this method. Hard to scale to 100m molecules.

Data preprocessing may have been prohibitive.



Docking

Almost no one attempted Docking methods.

Too expensive to run docking on full validation set.

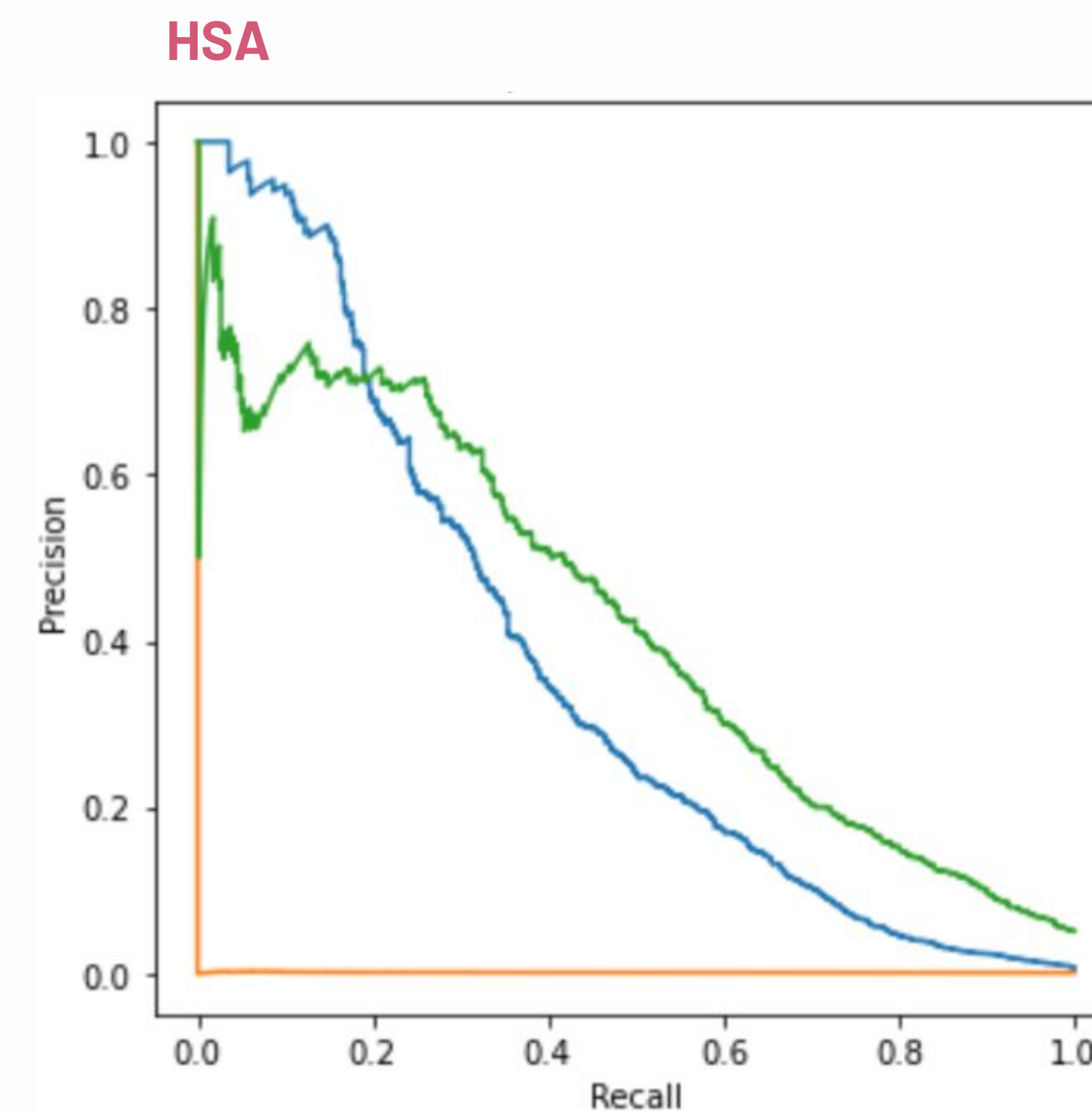
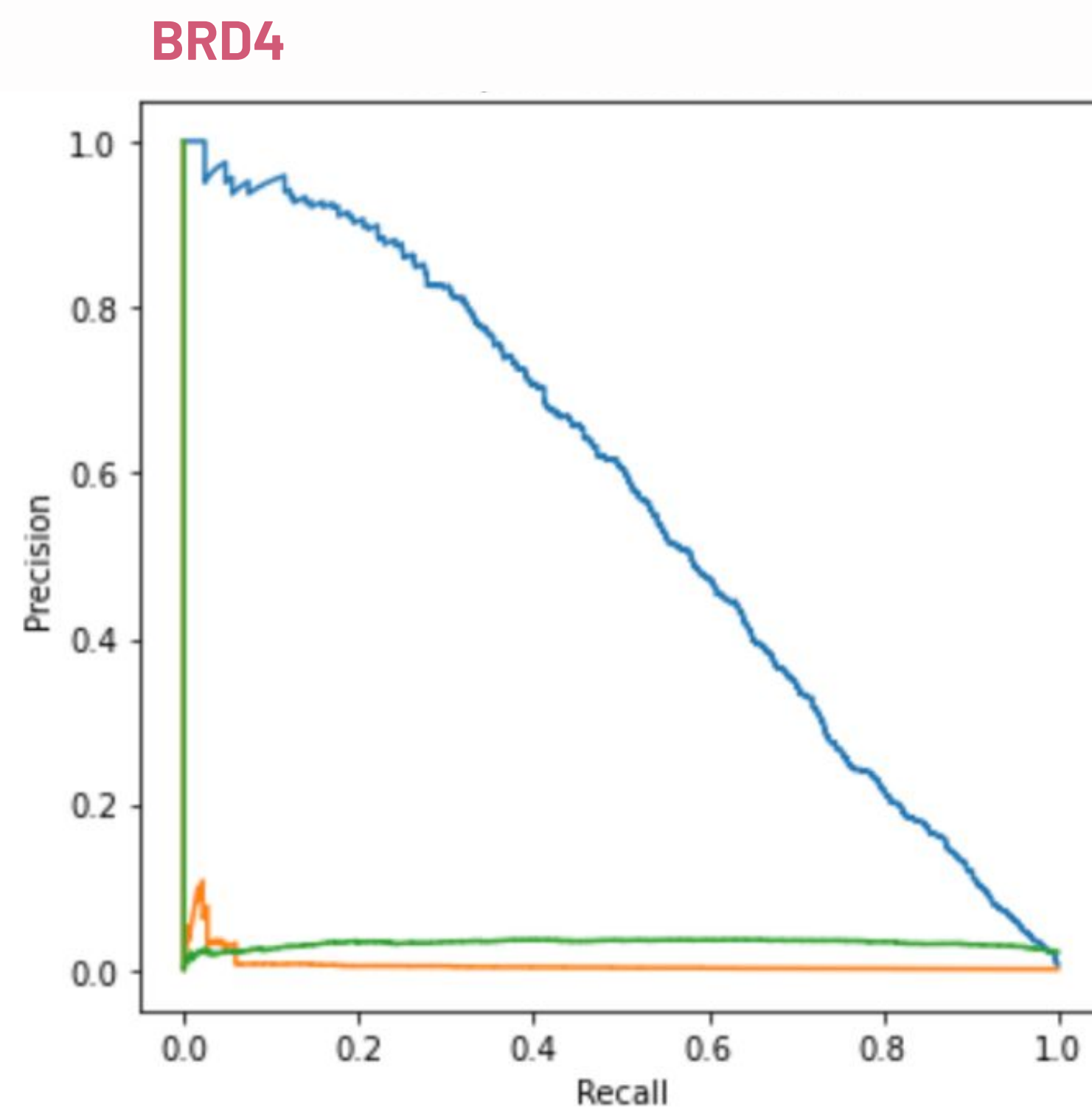
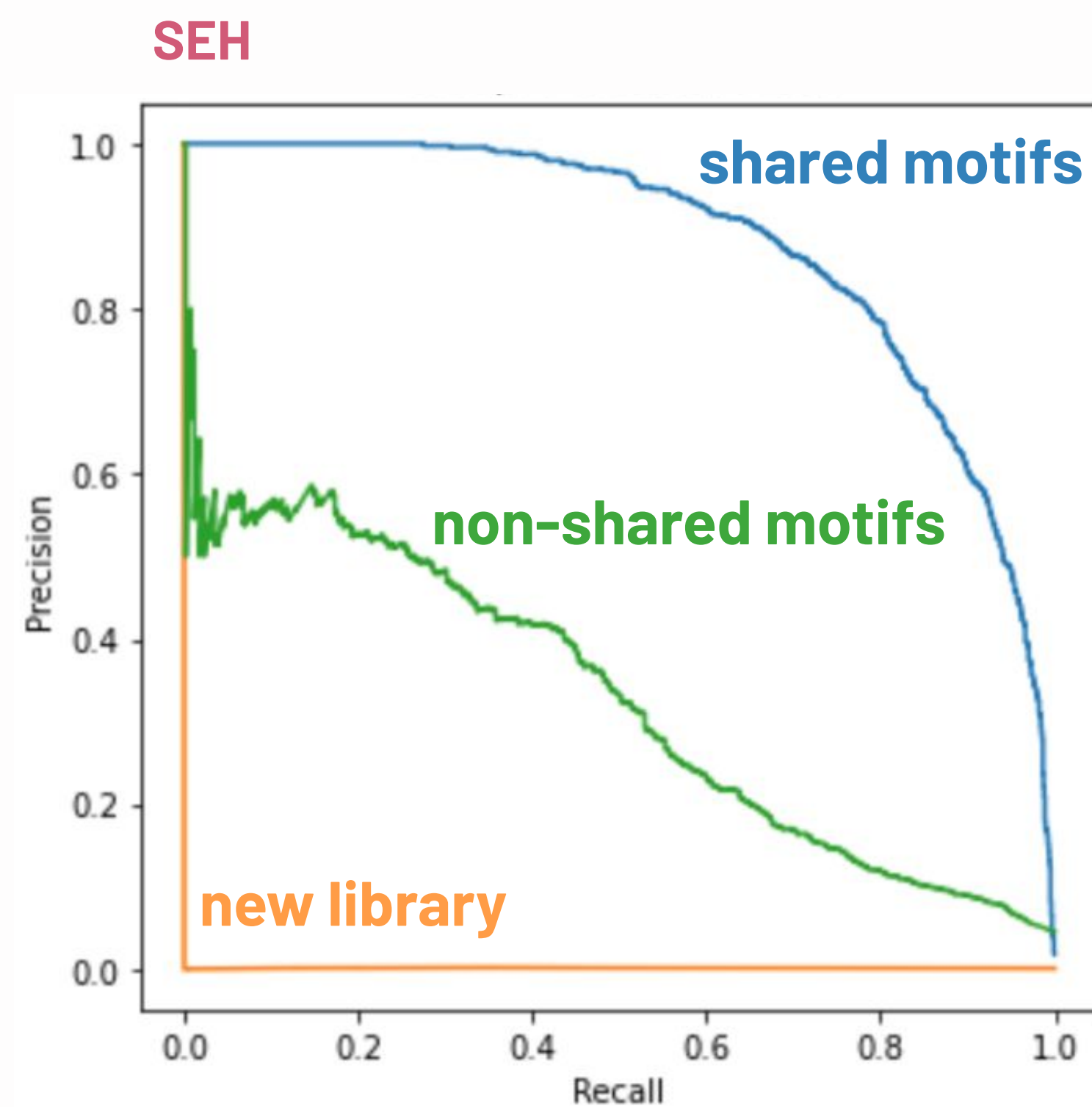
one group tried docking small sample and found no correlation between docking score and binding.



Kaggle results: participant's models mostly memorize

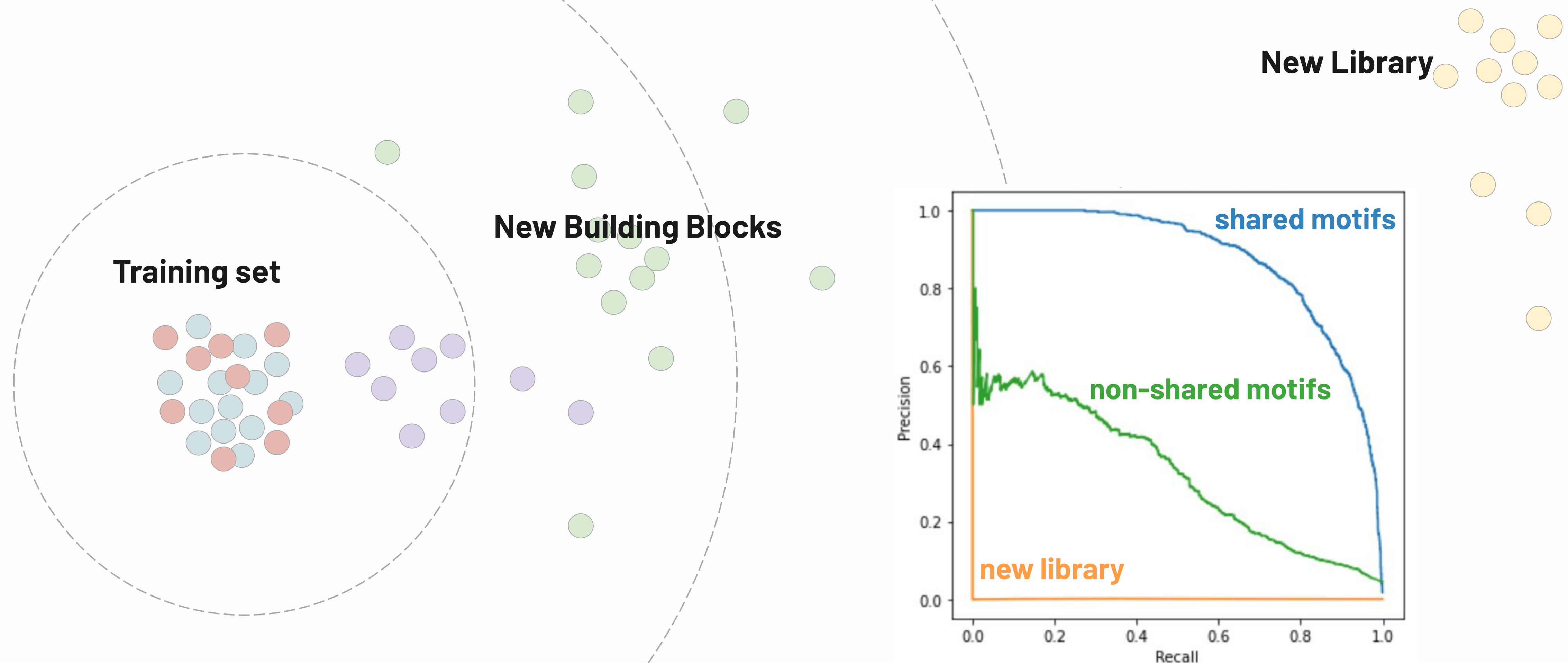
KAGGLE CONTESTANTS

- Do well on motifs shared in the training set
- Do less well on non-shared motifs
- Don't generalize at all to new libraries





BELKA split distributions





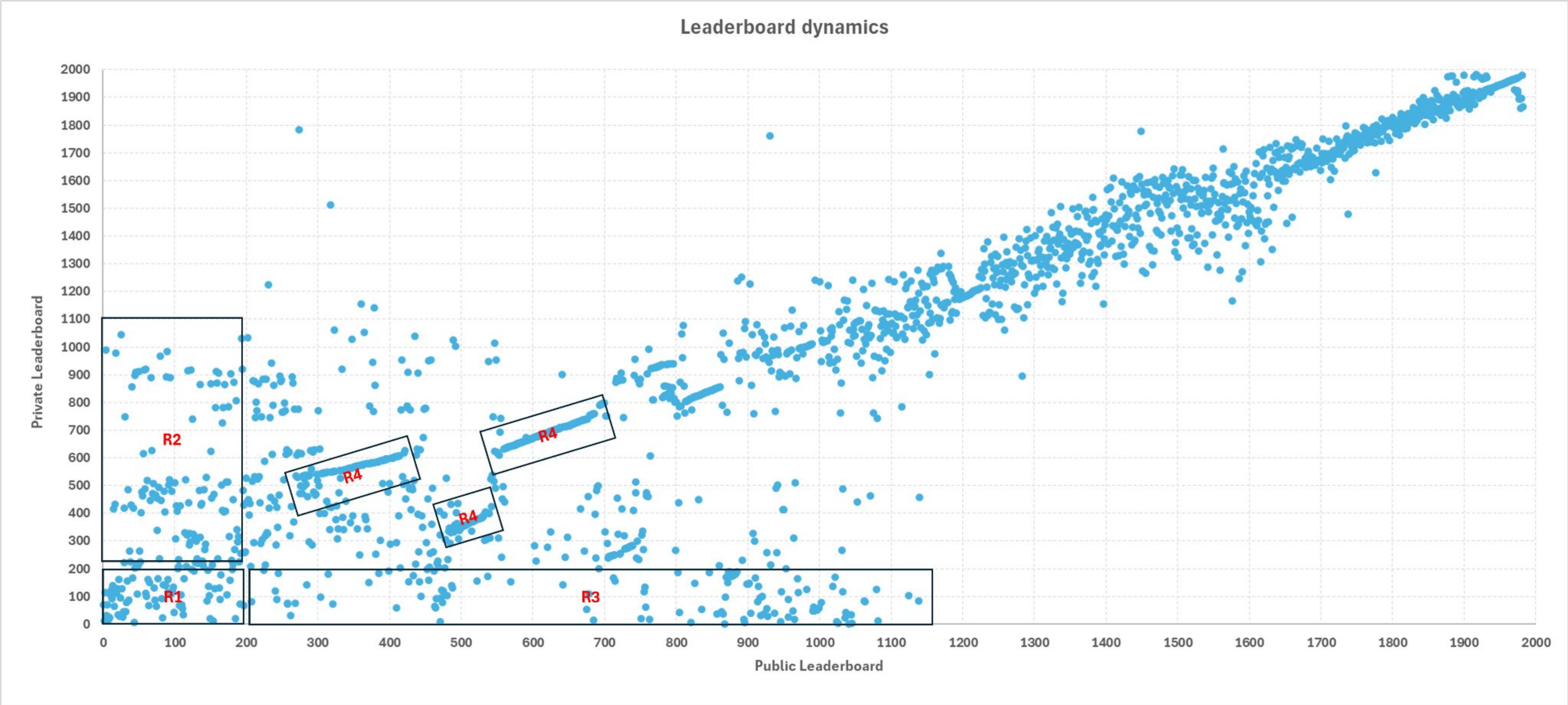
always has been

It's all similarity lookup



viewing the churn

Kaggle results: Shakeup





THANK YOU

belka@leash.bio