

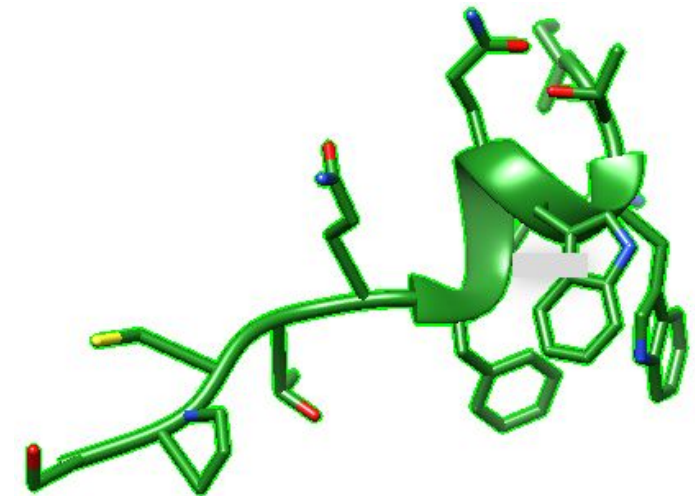
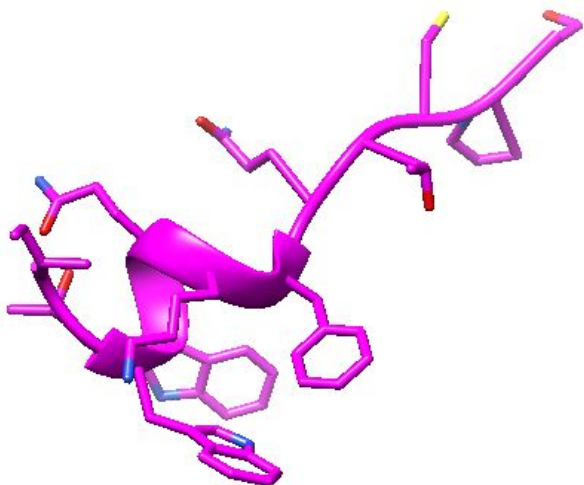
MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

Alp Deniz Ögüt

Biotechnology Graduate Programme

Masters' Defense

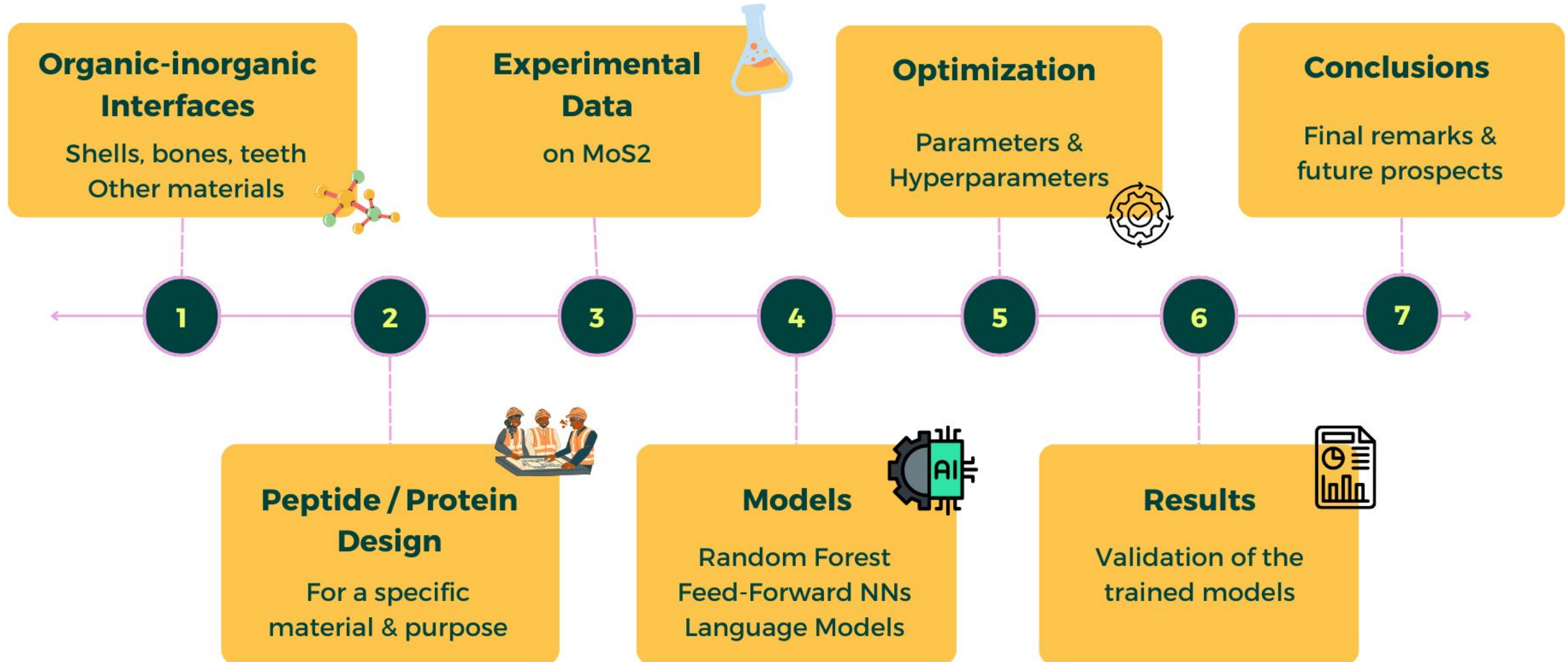
12 July 2024





MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

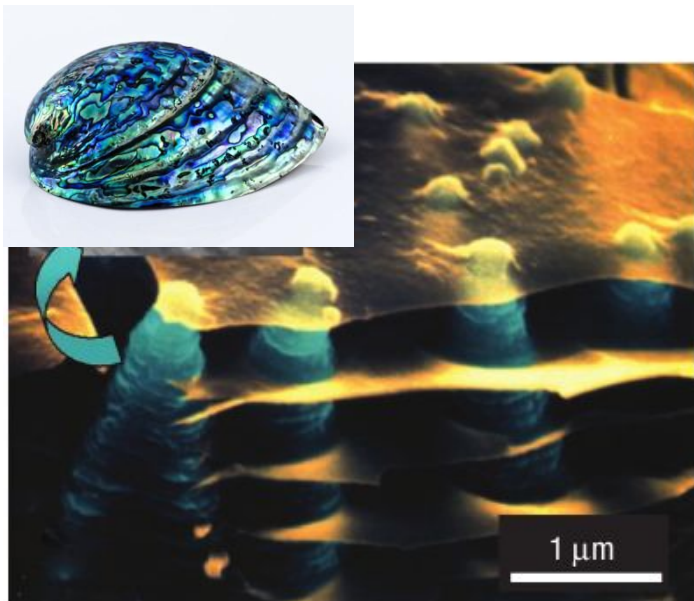
OUTLINE



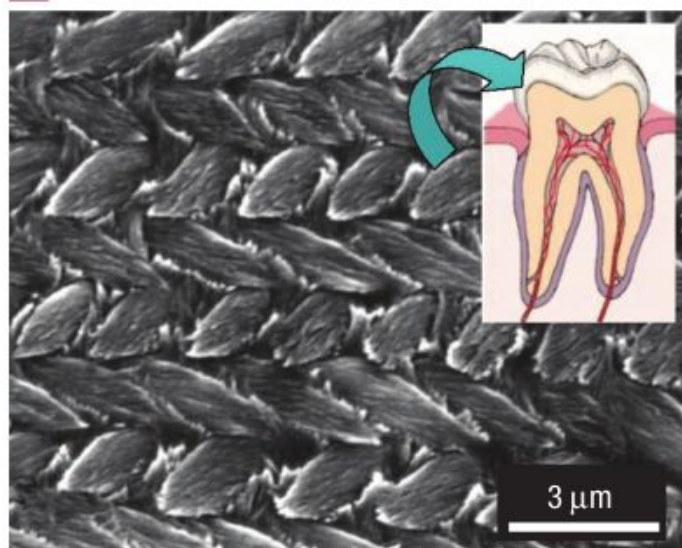


MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

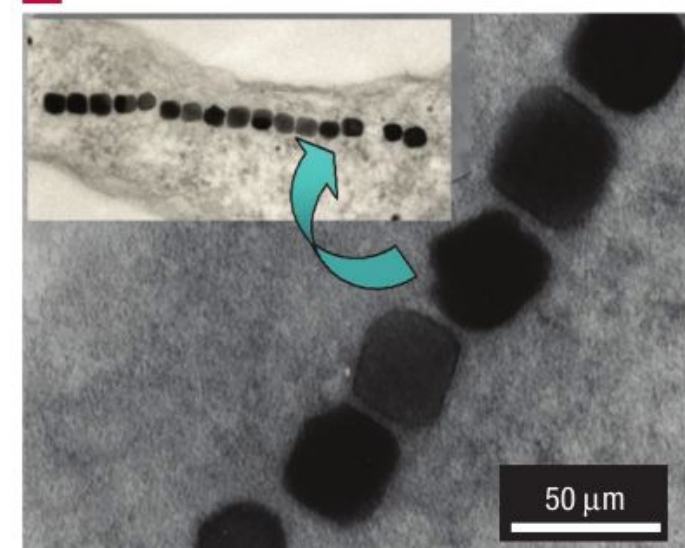
ORGANIC-INORGANIC INTERACTIONS



Growing inorganic structures and organic films (SEM)*



Mouse enamel (SEM)*



Magnetotactic bacterium and magnetite lumps (TEM)*

Evolved proteins interact with surrounding inorganic materials with high specificity

Specificity -> ions, solid materials

Function -> binding, biomineralization

*Sarıkaya et al. (2003). Molecular biomimetics: nanotechnology through biology. *Nature materials*, 2(9), 577-585.



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

PEPTIDE / PROTEIN DESIGN

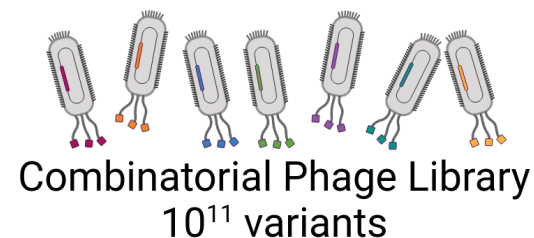
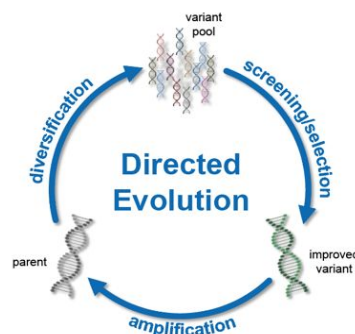
Recognizing atomic and molecular structures

Rational Design

- o Using known residue characteristics
- o No information about 3D conformation

Directed Evolution*

- o Mimics the natural evolution
- o Selection through a specified function
- o Low-throughput

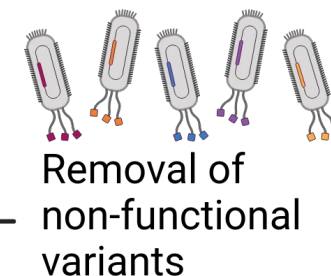


Selection towards
a specific function

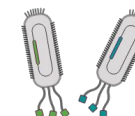
wash 1

wash 2

wash 3



Elution with targeted function
 10^4 variants
< 50 identified



*Arnold, F. H. (1998). Design by directed evolution. *Accounts of chemical research*, 31(3), 125-131.



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

DEEP-DIRECTED EVOLUTION

A significant improvement on Directed Evolution

Deep-Directed Evolution*

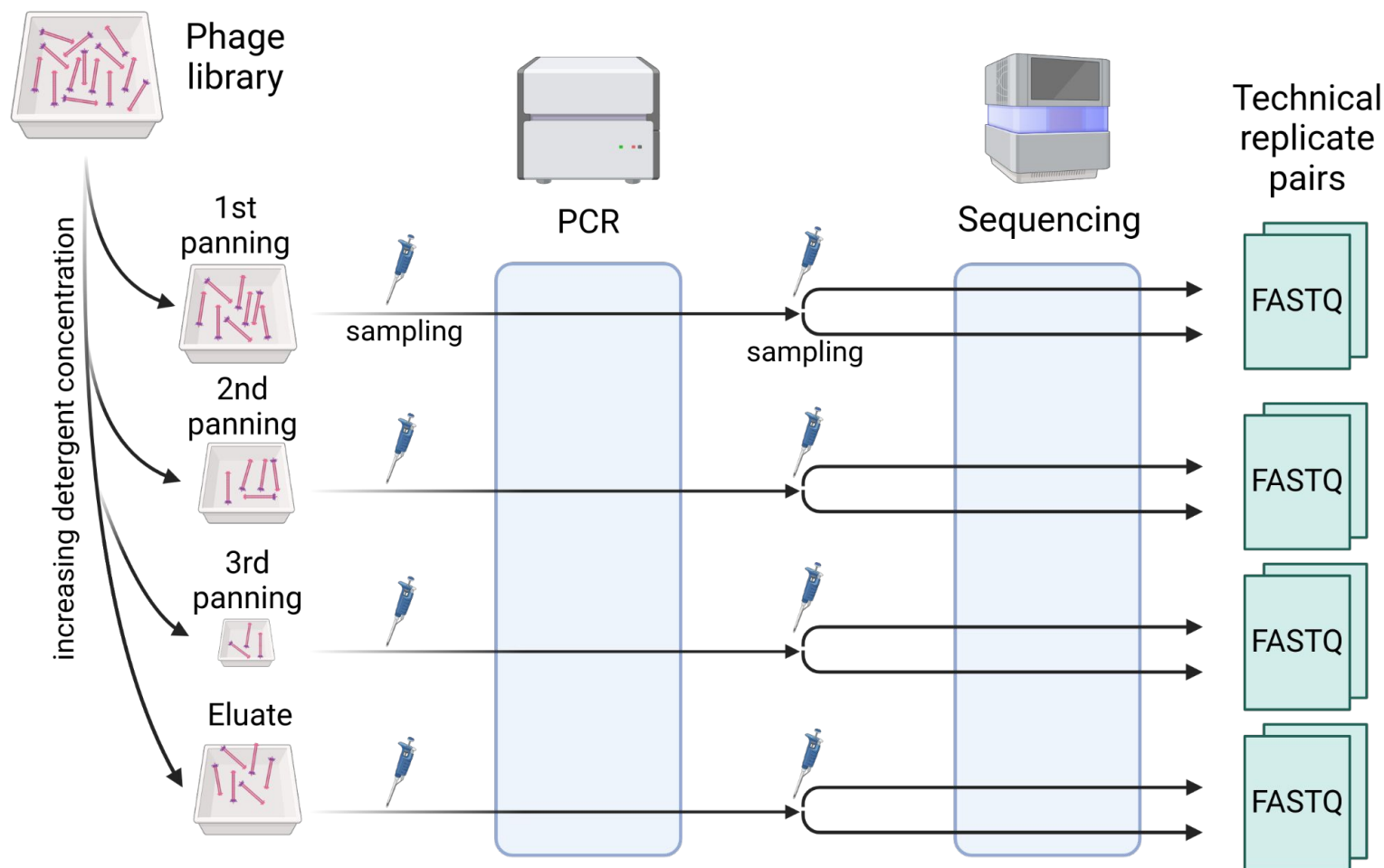
- Collects massive amounts of data
- Aims to capture sequence-function relationship

Achieved by:

- High-throughput NGS
- Advanced ML

Enables:

- Training machine learning models
- Searching the trained models





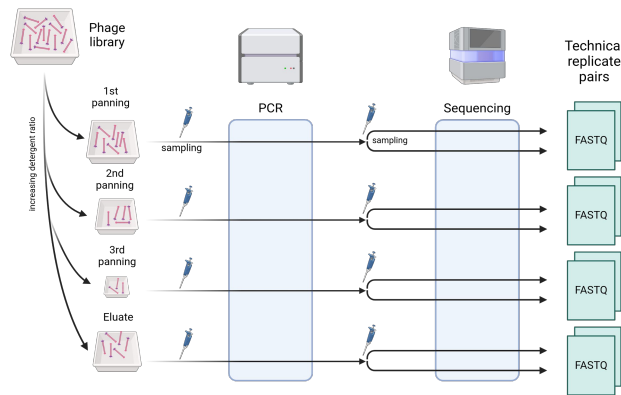
HYPOTHESIS and AIMS

Hypothesis:

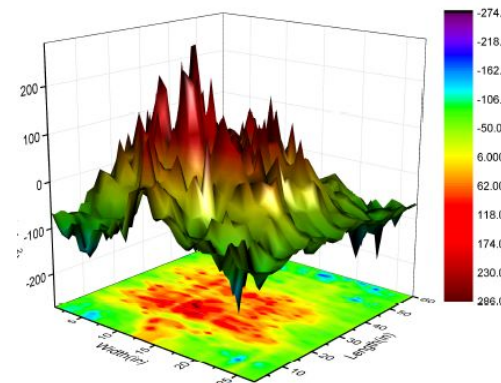
Deep-directed evolution sequencing data can be utilized to map out the entire sequence-function landscape.

Aims:

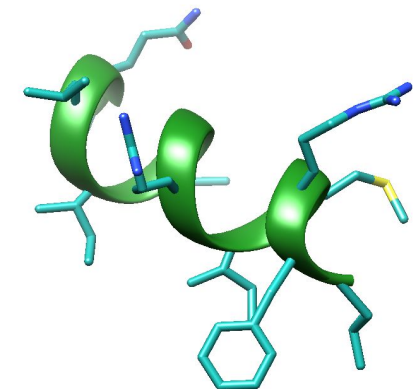
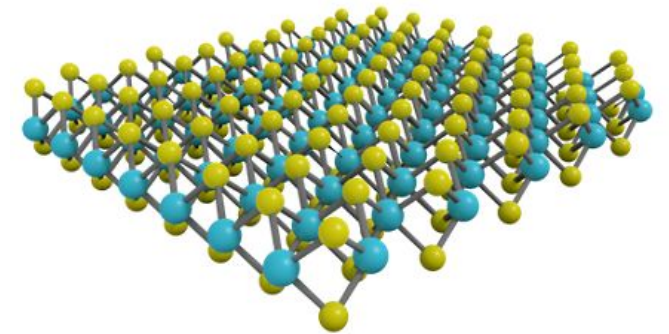
1. Process the experiment data
2. Model the sequence-function relationship with machine-learning methods.
3. Explore the model to design de novo peptides with the desired function



Experiment



Model



Functional peptide



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

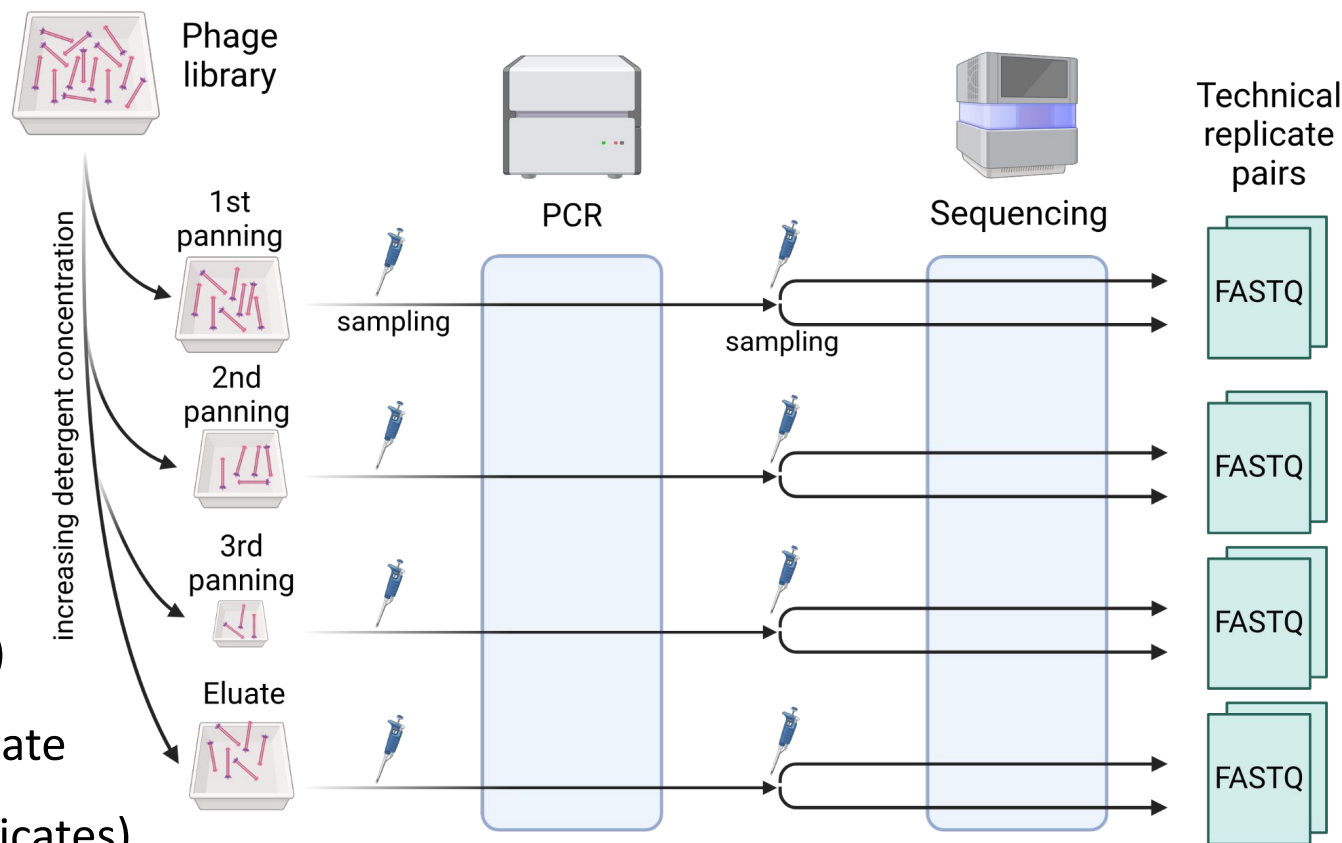
THE DATA

Deep-Directed Evolution Experiment*

- 24 FASTQ files
- 32GB of DNA sequences
- Each sequence is 36 nucleotides long

Files:

- 3 sets of parallel experiments (biological replicates)
- 4 panning steps: 1st wash, 2nd wash, 3rd wash, eluate
- 2 NGS runs for a single panning step (technical replicates)





MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

PREPROCESSING

Counting DNA sequences (set 1 - wash 2)

Sequence	count
AAAAAAAAAAAAAATGCCTTTGTGGATTCGGTCGAAT	5
AAAAAAAAAAAAAATATGCCATTGTGGATTCGGTCGAAA	2



Tabularization of sequences (set 1)

Sequence	wash1	wash2	wash3	eluate	total
AAAAAAAAAAAAAATGCCTTTGTGGATTCGGTCGAAT	14	5	0	7	26
AAAAAAAAAAAAAATATGCCATTGTGGATTCGGTCGAAA	0	2	0	2	4



Translation into peptides (set 1)

Peptide	wash1	wash2	wash3	eluate	total
KKKKMPLWIRSN	15	6	0	7	28
KKKNMPLWIRSK	3	7	0	1	11



# total DNA sequences	~222M
# unique DNA sequences	~44M
# unique amino acid sequences	~24M

Set	peptide	wash1	wash2	wash3	eluate	total
1	KKKKMPLWIRSN	15	6	0	7	28
2	KKKKMPLWIRSN	24	1	6	12	43
3	KKKKMPLWIRSN	20	0	0	14	34
Sum		59	7	6	33	105
1	KKKNMPLWIRSK	3	7	0	1	11
2	KKKNMPLWIRSK	12	0	0		
2	14					
3	KKKNMPLWIRSK	6	0	0	5	11
Sum		21	7	0	8	36



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

PREPROCESSING

Calculation of binding score

Assigning a score to a peptide:

- Measuring resistance to detergents
- In range: [0, 1]

Using the metric in the reference study:
Center of abundance-mass (CoAM)

$$\text{binding score} = \frac{1 * \text{wash2} + 2 * \text{wash3} + 3 * \text{eluate}}{3 * \text{total}}$$

Another sample from the dataset that demonstrates high binding affinity in all 3 biological sets

Set	peptide	wash1	wash2	wash3	eluate	total	score
1	VSWPWAWSRIQ	18	57	2	182	259	0.78
2	VSWPWAWSRIQ	31	24	0	139	194	0.76
3	VSWPWAWSRIQ	10	0	0	158	168	0.94
Sum		59	81	2	479	621	0.82

Merging of biological sets & calculation of binding scores

Set	peptide	wash1	wash2	wash3	eluate	total	score
1	KKKKMPLWIRSN	15	6	0	7	28	0.32
2	KKKKMPLWIRSN	24	1	6	12	43	0.38
3	KKKKMPLWIRSN	20	0	0	14	34	0.41
Sum		59	7	6	33	105	0.37
1	KKKNMPLWIRSK	3	7	0	1	11	0.30
2	KKKNMPLWIRSK	12	0	0	2	14	0.14
3	KKKNMPLWIRSK	6	0	0	5	11	0.45
Sum		21	7	0	8	36	0.28



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

PREPROCESSING

Total number of observations

Total count (“total” column):

- Total number of observations of a peptide across all 24 sequencing runs

Count filter N:

- Removes data rows of which total count is less than or equal to N
- Referred to as cfN
- Controls the noise/dataset-size trade off

Set	peptide	wash1	wash2	wash3	eluate	total	score
1	KKKKMPLWIRSN	15	6	0	7	28	0.32
2	KKKKMPLWIRSN	24	1	6	12	43	0.38
3	KKKKMPLWIRSN	20	0	0	14	34	0.41
Sum		59	7	6	33	105	0.37
1	KKKNMPLWIRSK	3	7	0	1	11	0.30
2	KKKNMPLWIRSK	12	0	0	2	14	0.14
3	KKKNMPLWIRSK	6	0	0	5	11	0.45
Sum		21	7	0	8	36	0.28
1	VSWPWAWSRIQ	18	57	2	182	259	0.78
2	VSWPWAWSRIQ	31	24	0	139	194	0.76
3	VSWPWAWSRIQ	10	0	0	158	168	0.94
Sum		59	81	2	479	621	0.82



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

CONFIDENCE of DATA POINTS

Sample weights

Total count:

- How confident are we about the score?

Key points:

- Law of large numbers
- Central limit theorem

Confidence of a peptide-score pair:

$$w_i = C_i^\varphi$$

where φ is chosen as 0.7

Set	peptide	wash1	wash2	wash3	eluate	total	score
1	KKKKMPLWIRSN	15	6	0	7	28	0.32
2	KKKKMPLWIRSN	24	1	6	12	43	0.38
3	KKKKMPLWIRSN	20	0	0	14	34	0.41
Sum		59	7	6	33	105	0.37
1	KKKNMPLWIRSK	3	7	0	1	11	0.30
2	KKKNMPLWIRSK	12	0	0	2	14	0.14
3	KKKNMPLWIRSK	6	0	0	5	11	0.45
Sum		21	7	0	8	36	0.28
1	VSWPWAWSRIQ	18	57	2	182	259	0.78
2	VSWPWAWSRIQ	31	24	0	139	194	0.76
3	VSWPWAWSRIQ	10	0	0	158	168	0.94
Sum		59	81	2	479	621	0.82

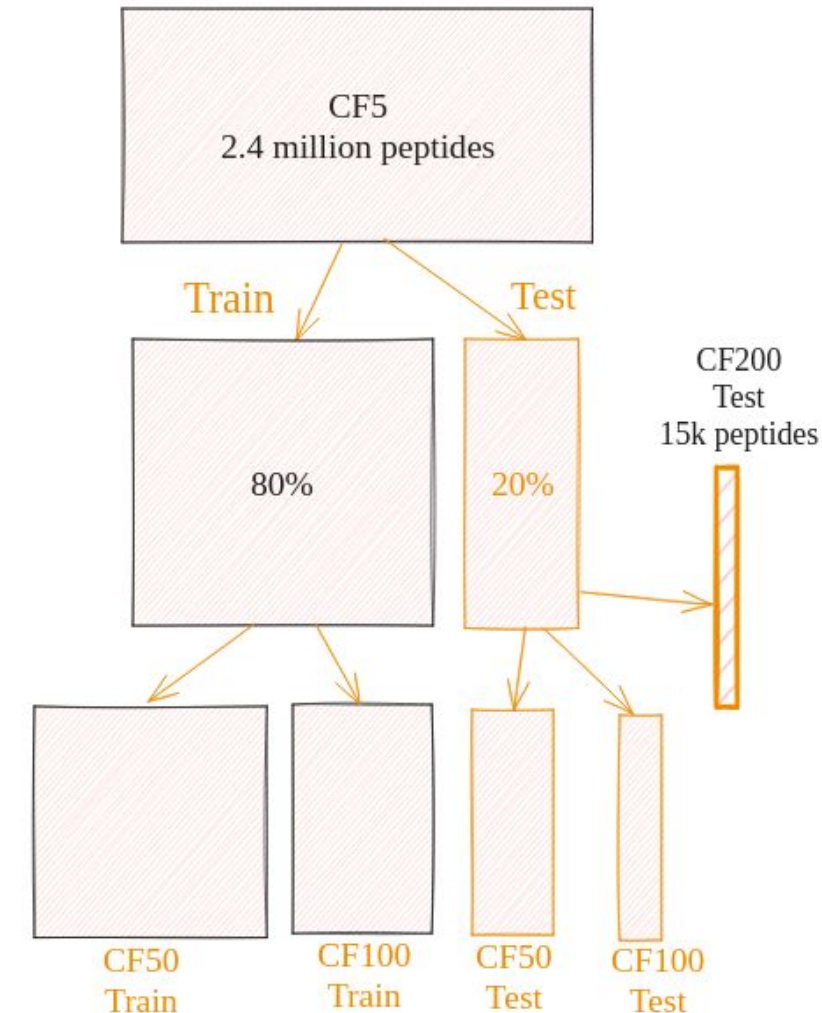
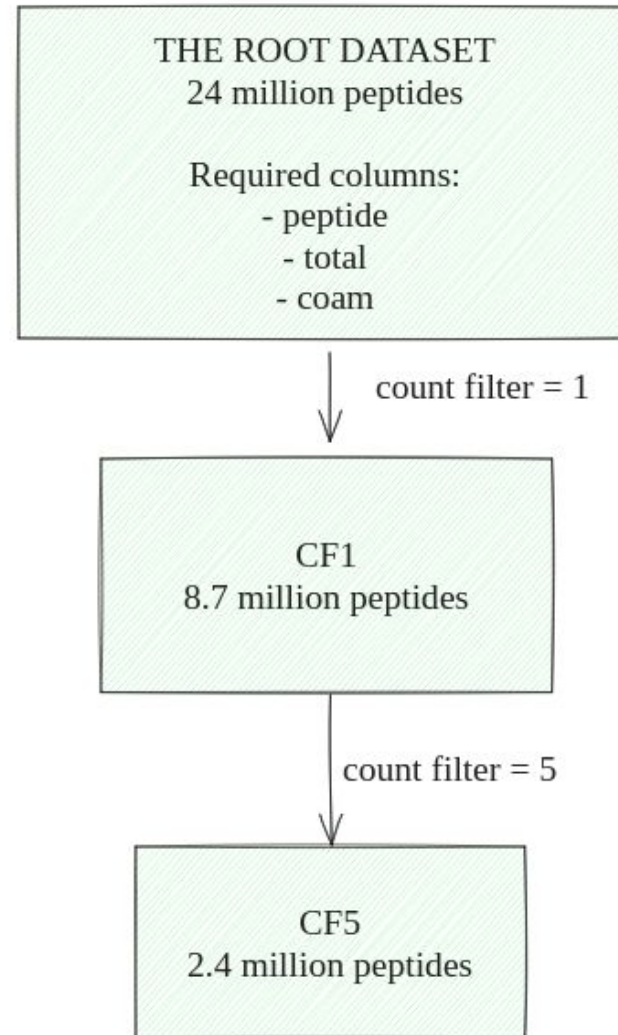


MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

PREPROCESSING

- Removed peptides with only one total number of observations to weed out potential sequencing errors
- Removed statistically insignificant peptides by using the count filter (cf5) as a starting point
- Train & Test Splits
 - Test set: 20%
 - Training: 80%

In case of further filtering, count filter is applied after the initial train/test split as shown.





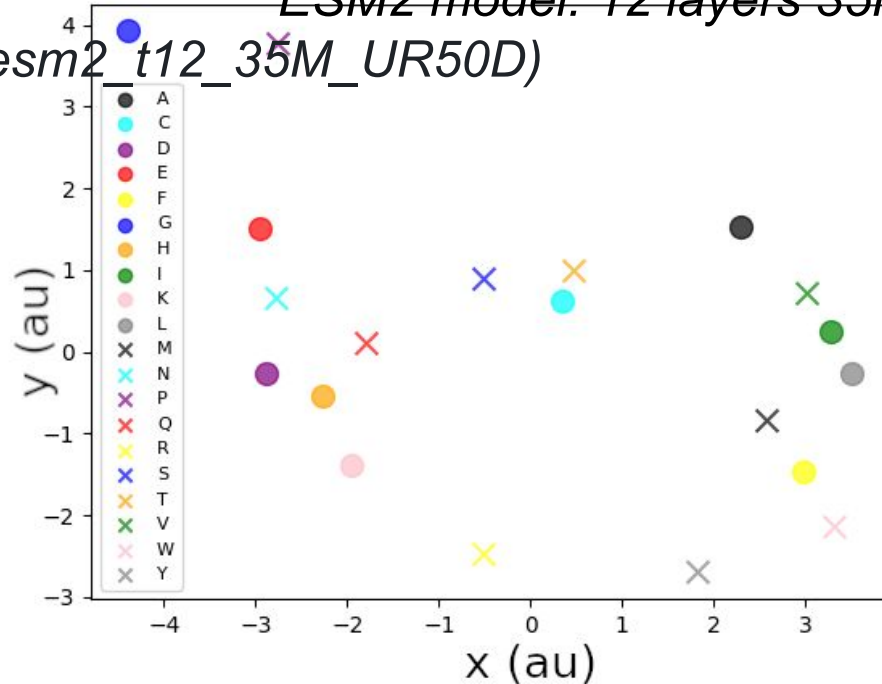
AMINO ACID ENCODINGS

One-hot - Residues represented as orthogonal binary vectors (20D)

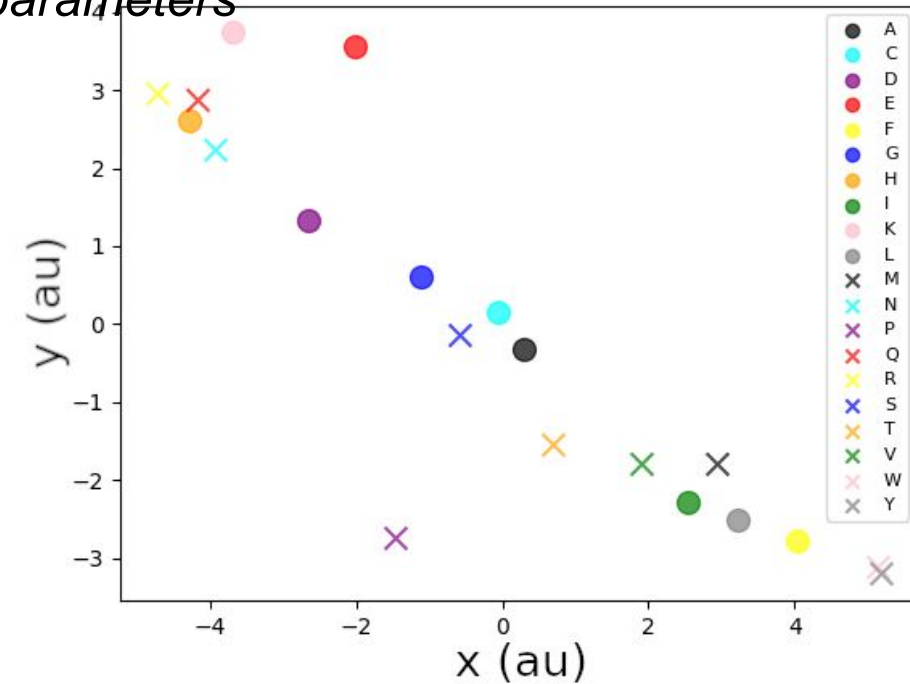
VHSE8 - Reduced hydrophobic, steric, and electronic properties (8D)*

ESM480 - Embedding weights of ESM Protein Language Model*
(480D)**

**ESM2 model: 12 layers 35M parameters*
(esm2_t12_35M_UR50D)



VHSE8 encoding vector of each amino acid, as projected with PaCMAP.



ESM480 encoding vector of each amino acid, as projected with PaCMAP.

* Mei, H. U. et al. (2005). A new set of amino acid descriptors and its application in peptide QSARs. *Peptide Science: Original Research on Biomolecules*, 80(6), 775-786.

** Rives, A. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), e2016239118.



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

THE DATASET

Unfiltered data:

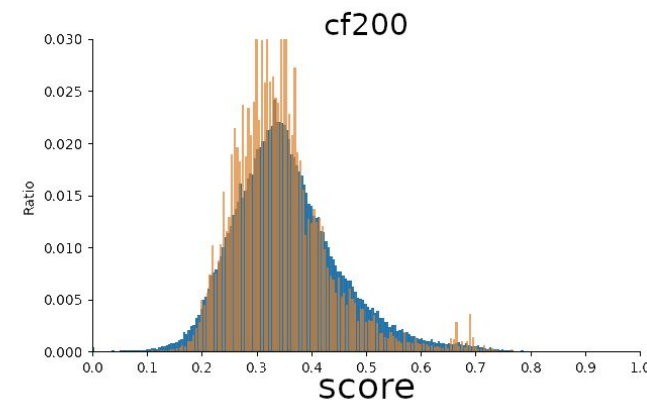
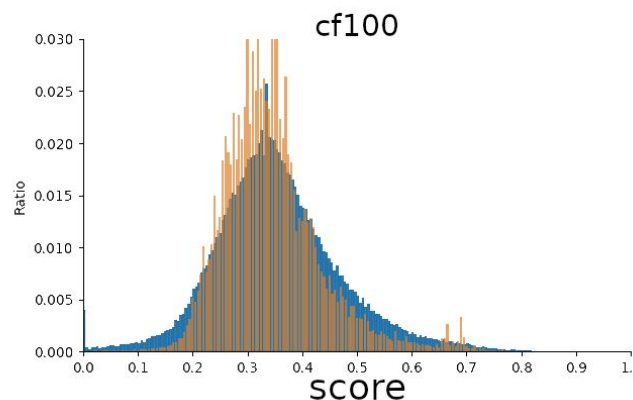
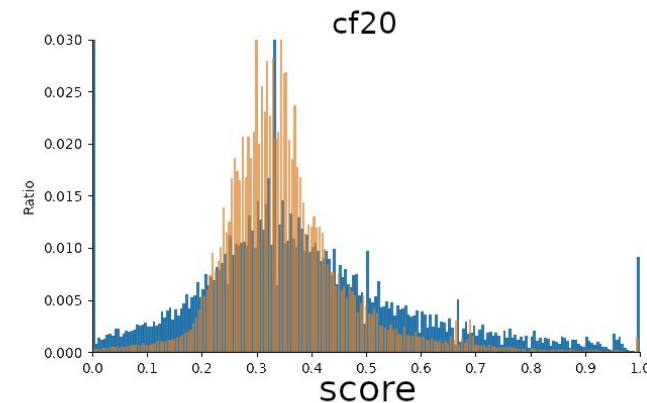
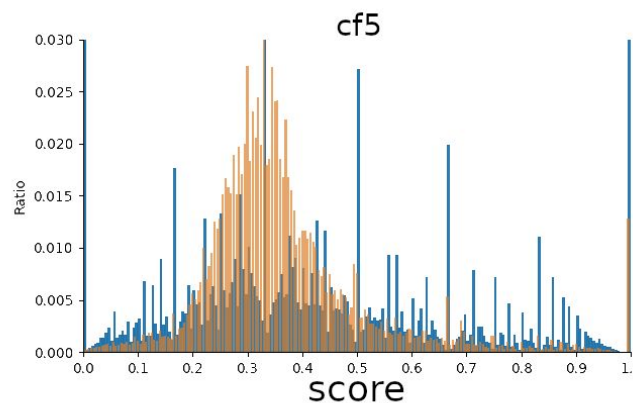
The distribution is closer to random distribution

Increasing count filter value:

The distribution gets closer to gaussian distribution

Observations:

- Peptides and their total observation counts are co-centered around the mean binding score (~0.35).
- Higher count filters suppresses the outliers.
- Dominant signal is around the mean.

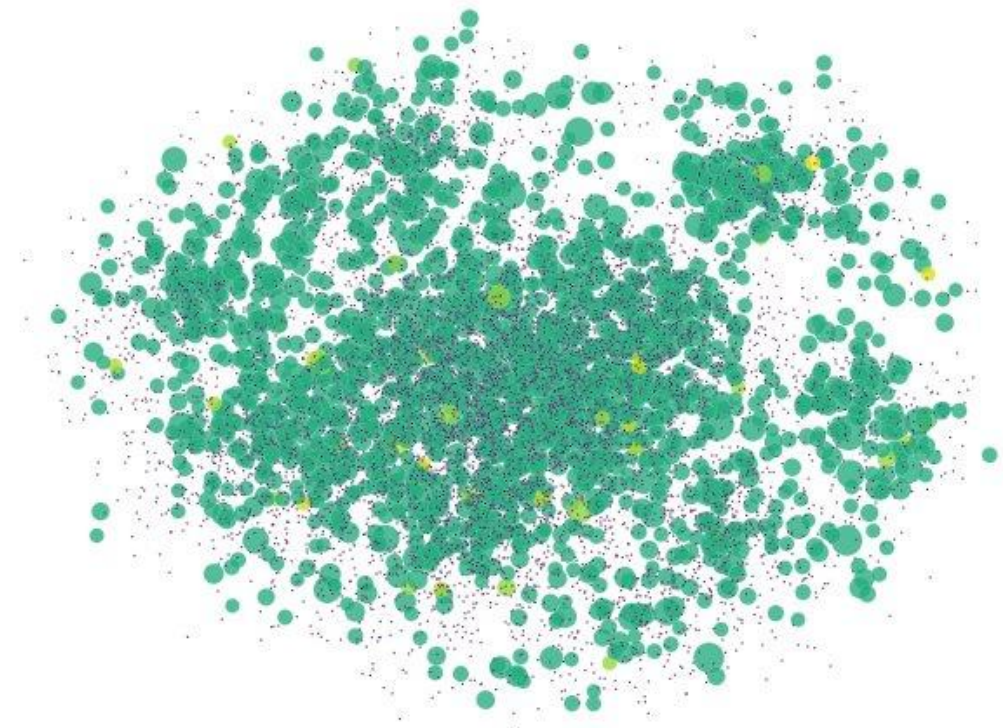


Normalized distribution of unique peptides (blue) and corresponding normalized total observation count distribution across the binding score range (orange).



THE DATASET

Phage library sequence space coverage seems adequate with no distant groups of random peptides.

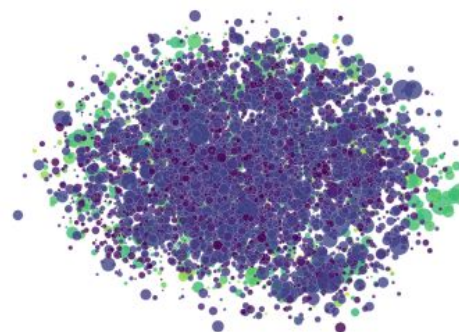


PacMap visualization of sequence space coverage.

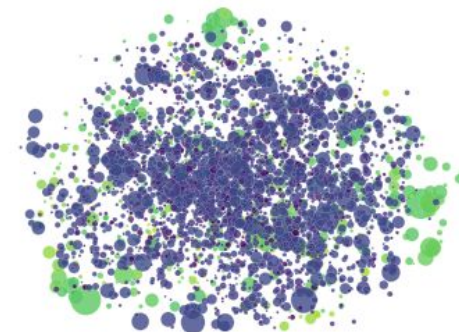
Sampled subset of cf200 set (green hues) against randomly generated peptides (dark blue) using VHSE8 encoding, where lighter green spots indicate strong binders.

At higher count filter values, the strong and weak binder clusters can be visually identified.

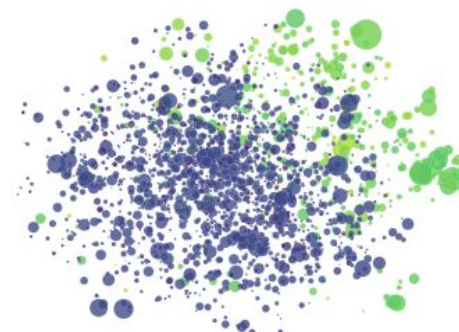
Dimensionality Reduction and Visual Clustering of VHSE8 Encoded Peptides



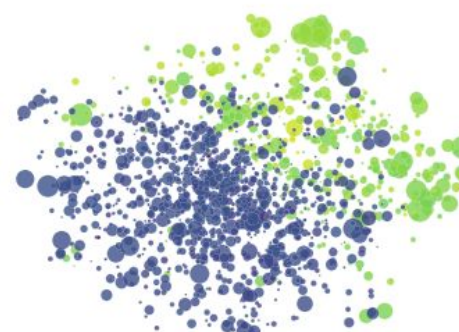
cf50



cf100



cf200



cf400

PacMap visualization of the dataset with VHSE8 encoding, using count filters 100, 150, 200, and 400.



RANDOM FOREST MODEL

Random Forest Models (RF)

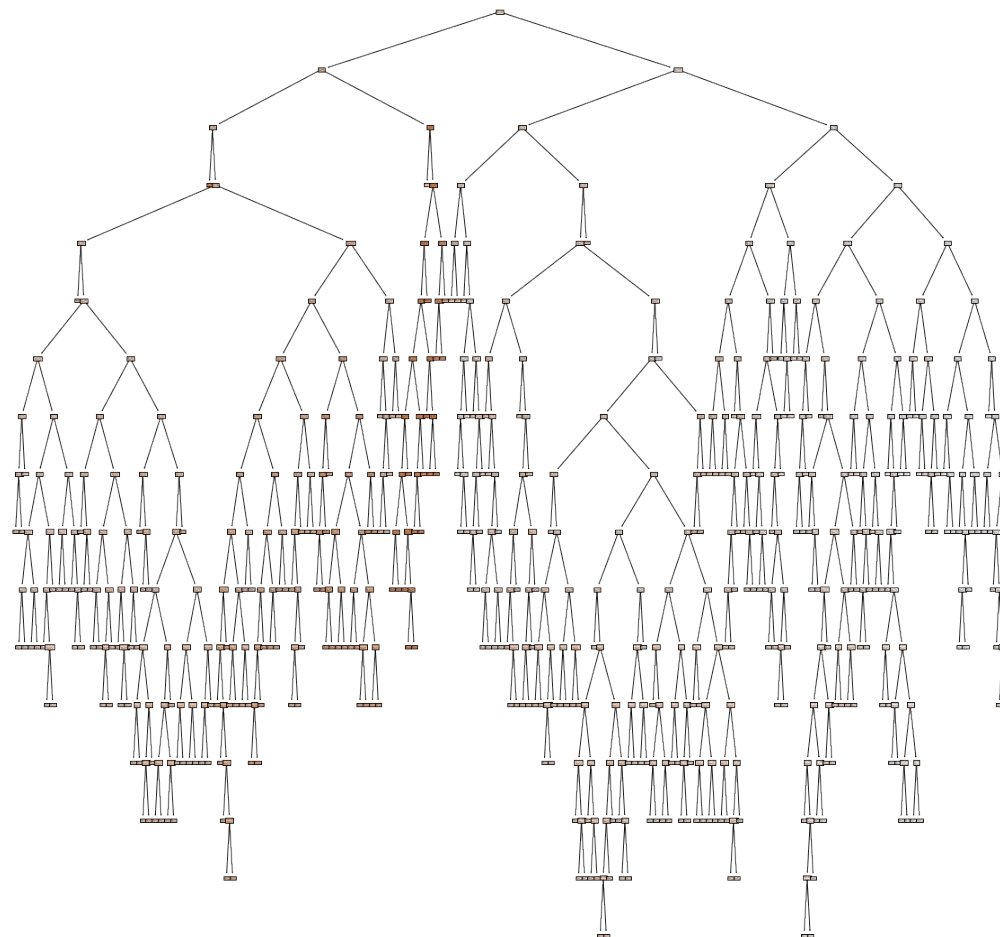
- Multiple decision trees
- Randomized subsampling

Pros:

- + Robust to noise
- + Few hyperparameters
- + Straightforward to fit & deploy

Cons:

- High memory / storage demand, proportional to data size, both for fitting and inference.





MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

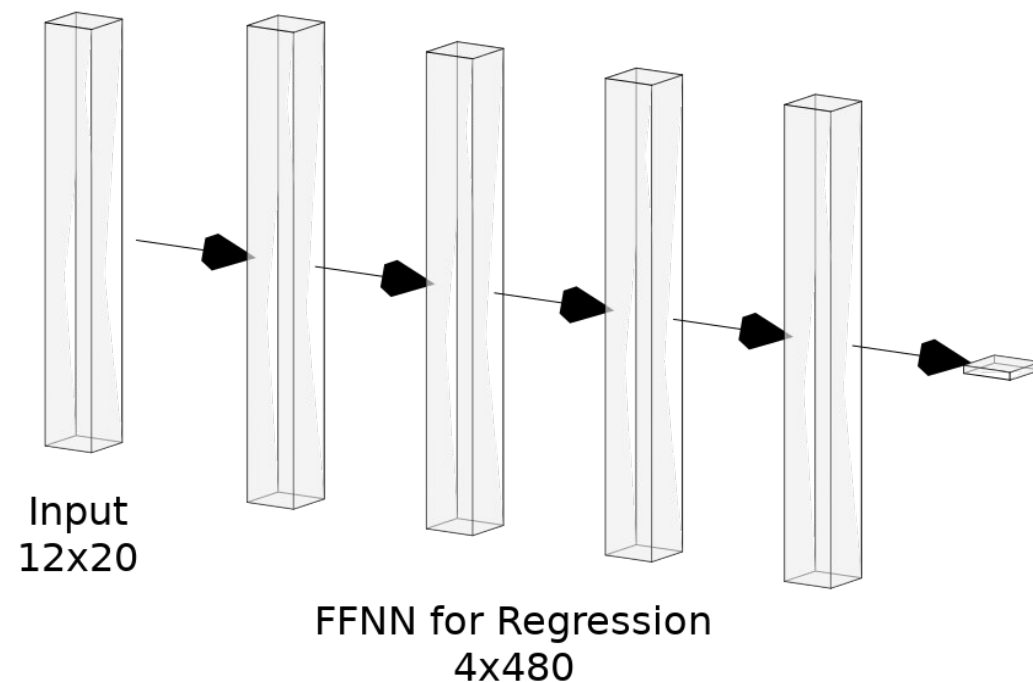
FEED-FORWARD NEURAL NETWORK MODEL

Feed-Forward Neural Network Models

Given an vectorized peptide
Predicts the binding score (CoAM)

Model:

- Input shape: 240 or 96 (one-hot and VHSE8)
- Output shape: 1
- 1, 4, 8 x layers
- 480 neurons per layer





Next token prediction: A simple unsupervised training objective

- Bengio et al. introduced neural network language models (2003)
- Vaswani et al. introduced the Transformer architecture (2017)
- And what does it mean to predict the next token well enough?
- The resulting vector is a unique representation of the input sequence as a whole.
- Learning causality!



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

PEPTIDE LANGUAGE MODEL

Peptide Language Models

Next token prediction:

Autoregressive unsupervised training

Given N amino acids, predicts the next one

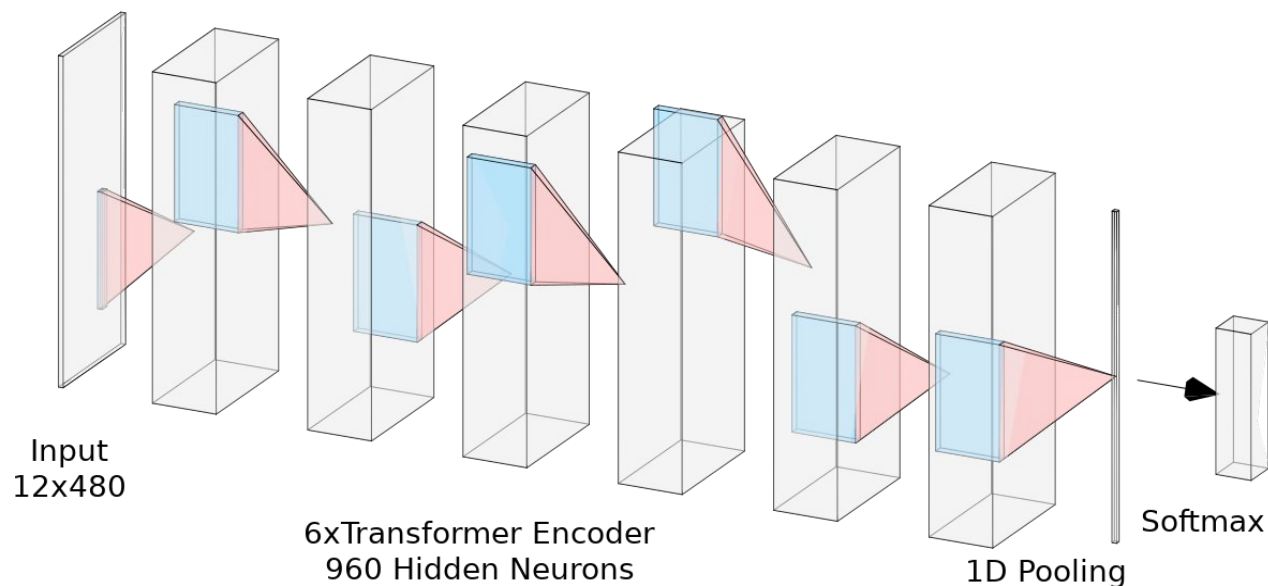
- Sequence length: 12
- Context: 2 up to 11

Encoding

- ESM480 encoding

Model:

- Input shape: 12 x 480 (ESM)
- Output shape: 20 residues
- 6 x Transformer encoder blocks
- 960 hidden units





MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

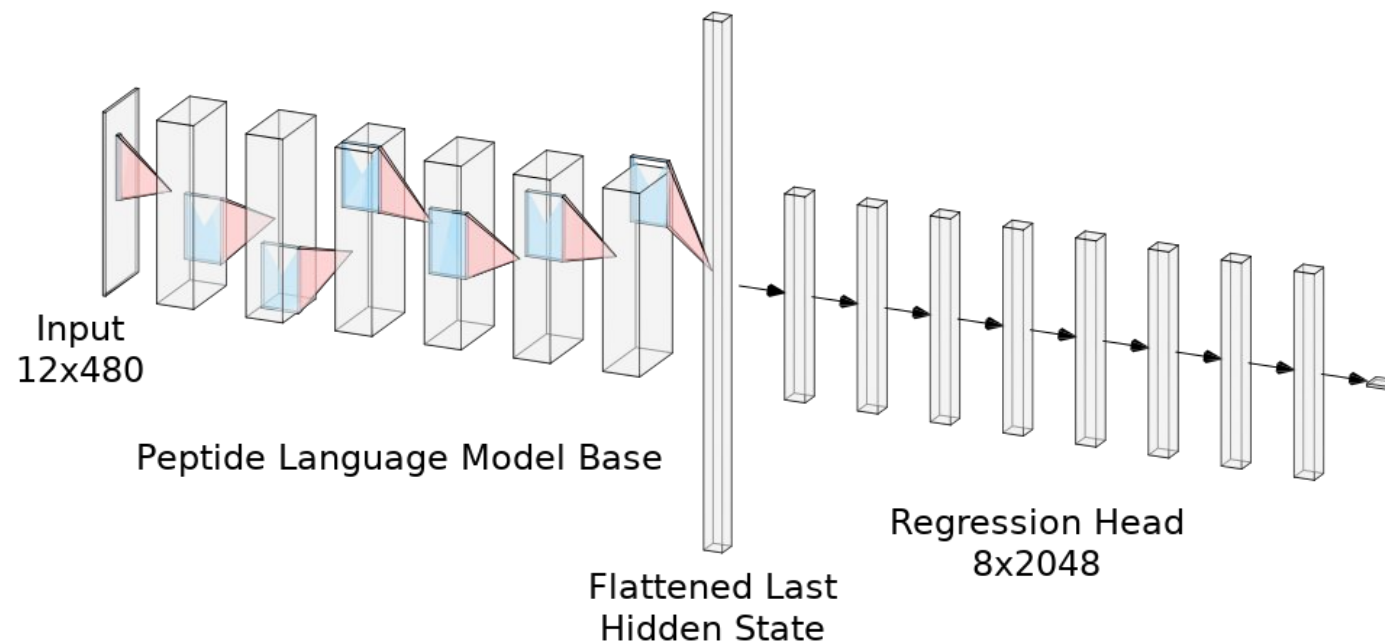
COMBINED LANGUAGE and REGRESSION MODEL

PepLM + FFNN Regression

Given an vectorized peptide
Predicts the binding score (CoAM)

Model:

- Input shape: 12 x 480 (ESM)
- Output shape: 1
- PepLM + 1, 8 x layers
- 480 and 2048 neurons per layer





MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

EFFECTS of COUNT FILTER on RF BASELINE PERFORMANCE

Training set	cf5	cf10	cf20	cf50	cf100	cf200
Training Data Size (rows)	2.1M	1.3M	696k	261k	120k	73k
Training Loss	0.0100	0.0084	0.0054	0.0025	0.0012	0.0007
Validation Loss	0.072	0.060	0.039	0.017	0.009	0.005
Test cf200 loss	0.0028	0.0029	0.0029	0.0037	0.0043	0.0049
Test cf200 Pearson	0.86	0.85	0.84	0.79	0.75	0.71
Model Storage Size (GB)	16.3	9.2	4.9	1.9	0.873	0.427

Different count filters resulted in

Comparing the effects of count filter on RF model prediction performance.



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

EFFECTS of SAMPLE WEIGHTS AND ENCODING

Transformed total number of observations are utilized as sample weights.

VHSE8 encodings are tested against uninformed one-hot representation.

Experiments demonstrate effectiveness of both methods:

- Sample weights

	One-hot	
Sample Weights	No	Yes
Training Loss	0.048	0.052
Validation Loss	0.063	0.063
Test 200 loss	0.0041	0.0034
Test 200 Pearson	0.79	0.82

	One-hot		VHSE8	
	Unweighted	Weighted	Unweighted	Weighted
Training Loss	0.048	0.052	0.032	0.037
Validation Loss	0.063	0.063	0.061	0.061
Test 200 loss	0.0041	0.0034	0.0032	0.0027
Test 200 Pearson	0.79	0.82	0.842	0.86



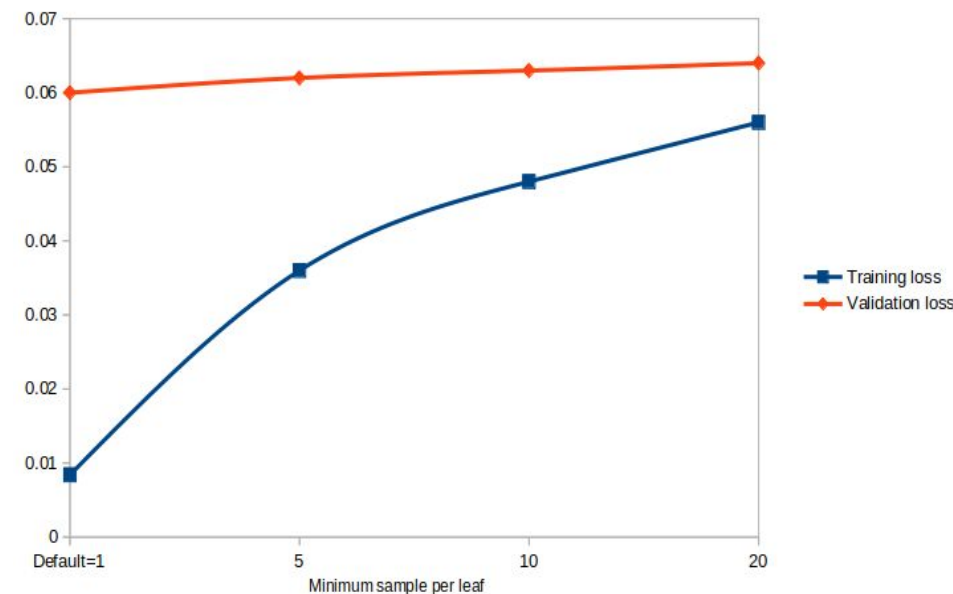
MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

RANDOM FOREST OPTIMIZATION

Random Forest Hyperparameters

- Dataset: cf5
- Number of estimators (decision trees): 200
- Minimum samples per leaf: 5

	Number Of Estimators				
Min sample per leaf	20	50	100	200	300
5	0.0812	0.0795	0.0789	0.0786	0.0785
10	0.0816	0.0805	0.0801	0.0798	0.0798
20	0.0822	0.0815	0.0813	0.0811	0.0811
50	0.0829	0.0826	0.0825	0.0824	0.0824
100	0.0834	0.0832	0.0831	0.0831	0.0831



Top: Trajectories of training and validation loss on increasing *minimum sample per leaf* value on cf10 dataset.

Left: 5-Fold cross-validation results of hyperparameter grid search on cf5 dataset.

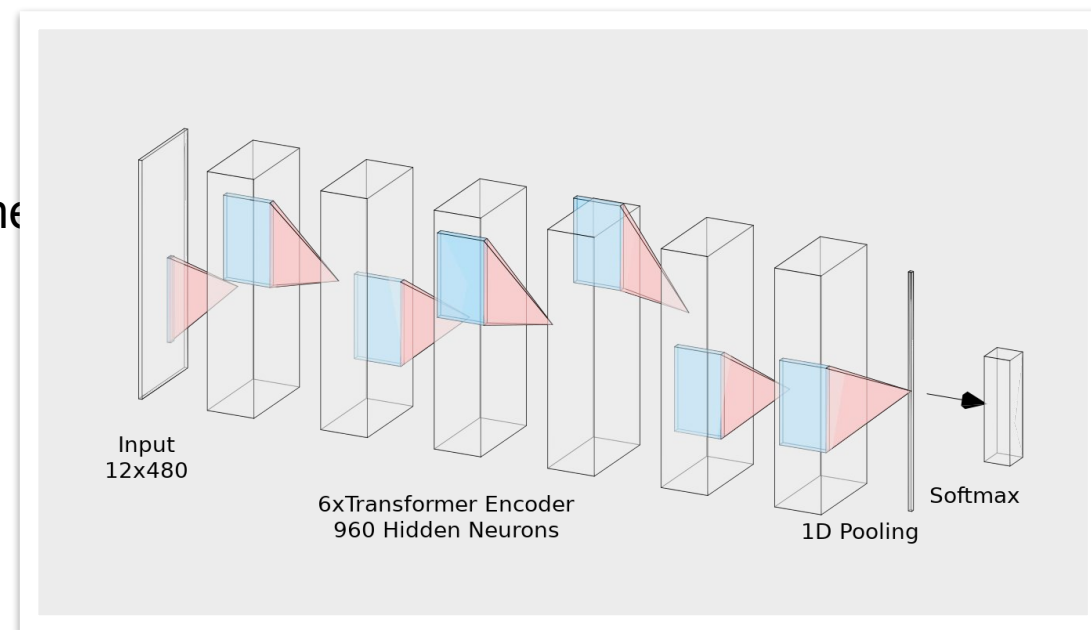


MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

LANGUAGE MODEL OPTIMIZATION

Peptide Language Models

- Largest model performs the best -> Room for improvement
- One-hot input model gets close to ESM input model
- Trained on 2M tokens
- For comparison
 - Google's PaLM achieves ~30% accuracy*
 - Facebook ESM-1b achieves ~28% accuracy**



Model	Train Accuracy	Test Accuracy
PepLM4 – ESM480 - 4xTransformer Layers	0.24	0.214
PepLM6 – Onehot - 6xTransformer Layers	0.253	0.228
PepLM6 – ESM480 - 6xTransformer Layers	0.251	0.232

*Chowdhery, A. et al. (2022). PaLM: Scaling language modeling with pathways.

**Rives, A. et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

NEURAL NETWORK OPTIMIZATION

Neural Network Hyperparameters

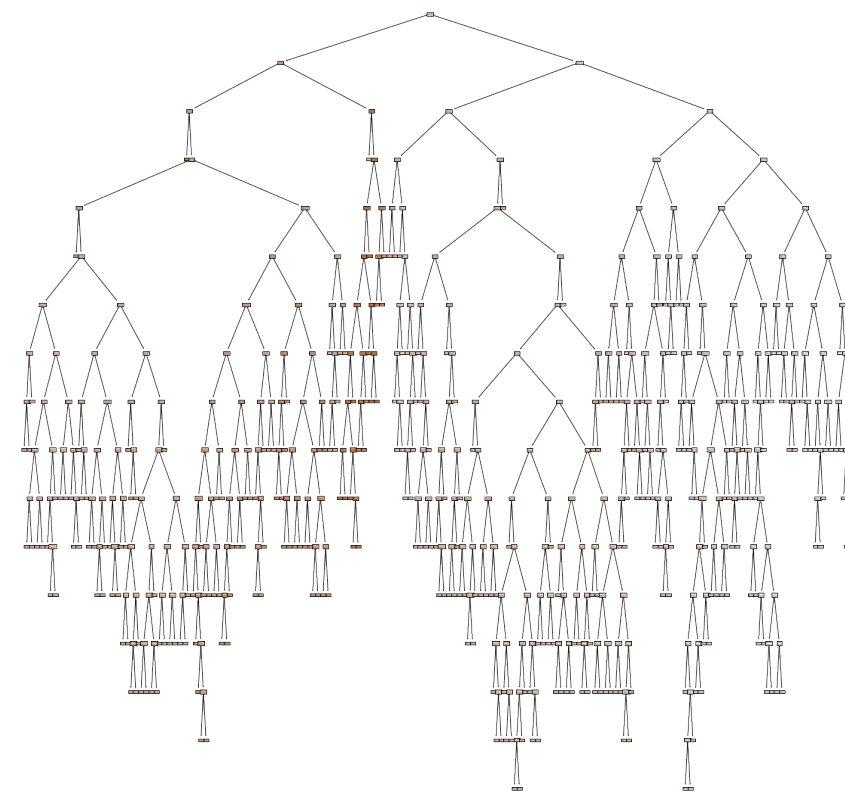
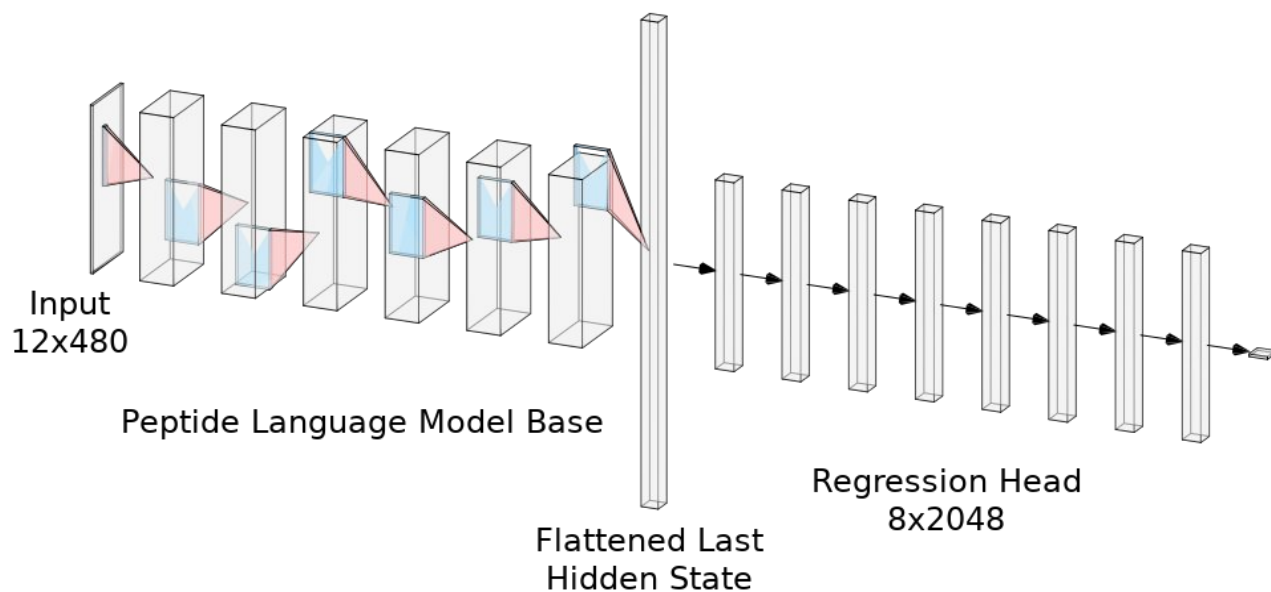
Explain the choices. Explain sampling methods (and data distribution)

Encoding	Model	Weighting	Sampling	Dropout	MSE	MAE
Onehot	FFNN 1x480	yes	no	0.5	0.0828	0.229
Onehot	FFNN 8x480	yes	no	0.5	0.0827	0.228
Onehot	FFNN 8x480	yes	oversampling	0.5	0.0847	0.238
Onehot	FFNN 8x480	no	oversampling	0.5	0.0927	0.256
VHSE8	FFNN 1x480	yes	no	0.5	0.0838	0.229
VHSE8	FFNN 8x480	yes	no	0.5	0.0828	0.227
ESM480	FFNN 1x480	yes	no	0.5	0.0824	0.227
ESM480	FFNN 8x480	yes	no	0.5	0.082	0.226
ESM480	PepLM+FFNN 1x480	yes	no	0.5	0.081	0.224
ESM480	PepLM+FFNN 8x480	yes	no	0.5	0.0802	0.221
ESM480	PepLM+FFNN 8x2048	yes	no	0.5	0.0757	0.21



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

RANDOM FOREST and NEURAL NETWORK ENSEMBLE MODEL



Averaging Ensemble

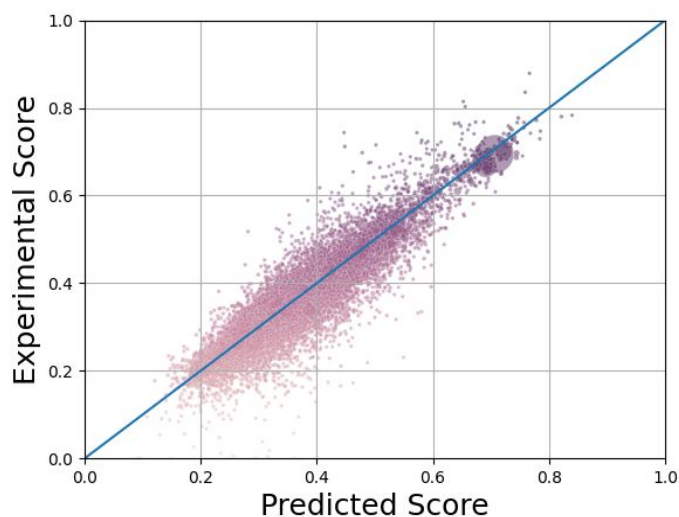


MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

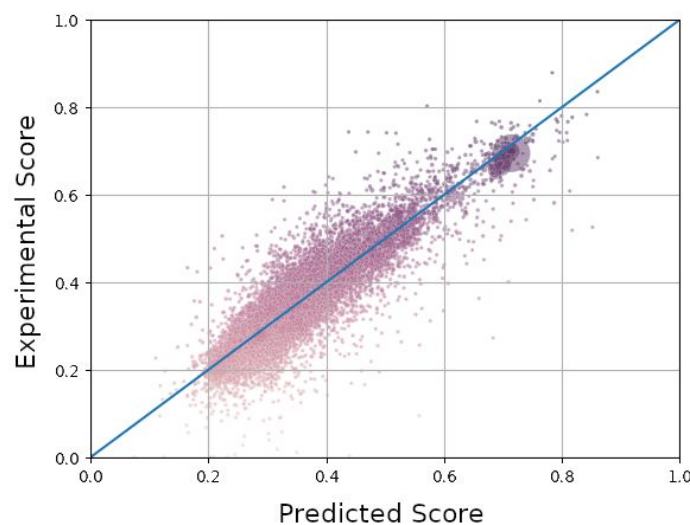
VALIDATION

Mean absolute error on high-confidence test set is around **3%** of the binding score range.

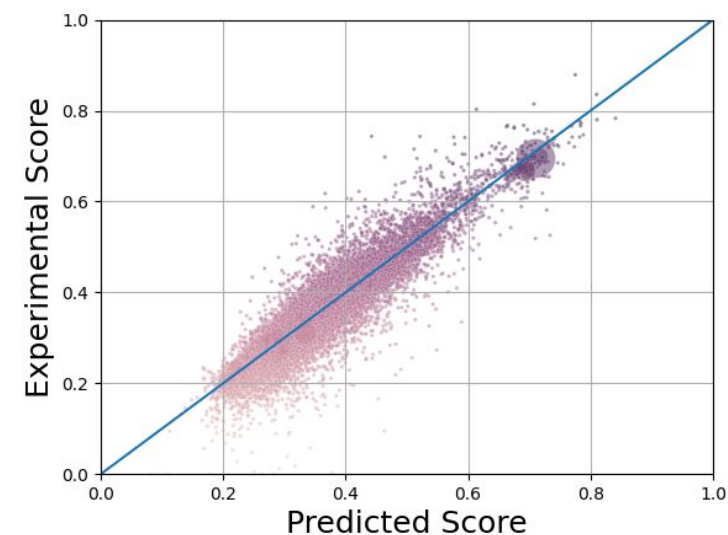
	RF	PepLM+FFNN	RF + NN Ensemble
Training cf5 Loss	0.0295	0.0413	0.0338
Test cf5 MSE Loss	0.0729	0.0746	0.0723
Test cf5 MAE	0.210	0.210	0.208
Test cf200 MSE Loss	0.00205	0.00242	0.00184
Test cf200 MAE	0.0325	0.0346	0.0304
Test cf200 Pearson	0.895	0.870	0.904



Optimized Random Forest



Optimized PepLM+FFNN



Ensemble Average



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

VALIDATION

Prediction results of the peptides collected from the literature

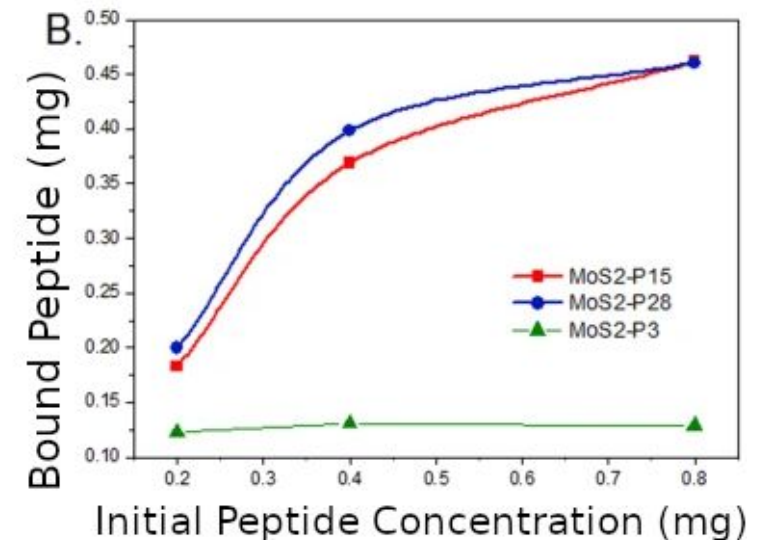
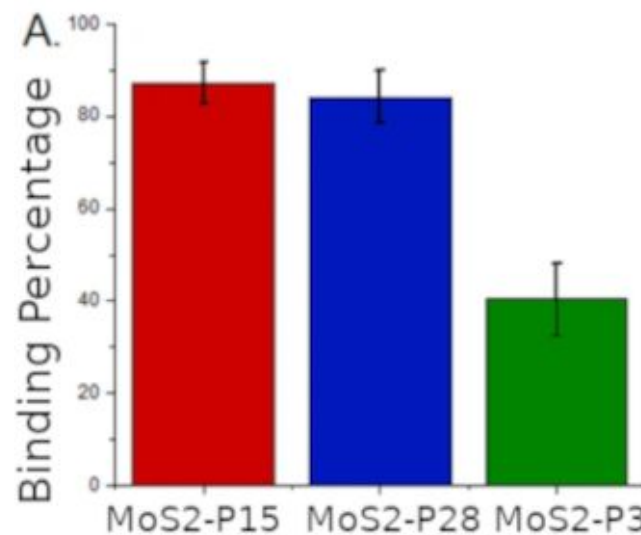
Peptide	Sequence	Affinity	RF Prediction	NN Prediction	Prediction
GrBP5-M6	IMVTASSAYDDY	Reference	0.41	0.32	0.36
MOS2-P15	GVIHRNDQWTAP	Strong	0.43	0.41	0.42
MOS2-P28	DRWVARDPASIF	Strong	0.44	0.39	0.41
MOS2-P3	SVMNTSTKDAIE	Weak	0.36	0.35	0.36

GrBP5-M6 is observed to weave nanowires

MOS2-PX peptides are experimentally compared:

- MOS2-P15 and P28 are strong binders
- MOS2-P3 is a weak binder.

Predictions are aligned with comparable results, however, prompts further





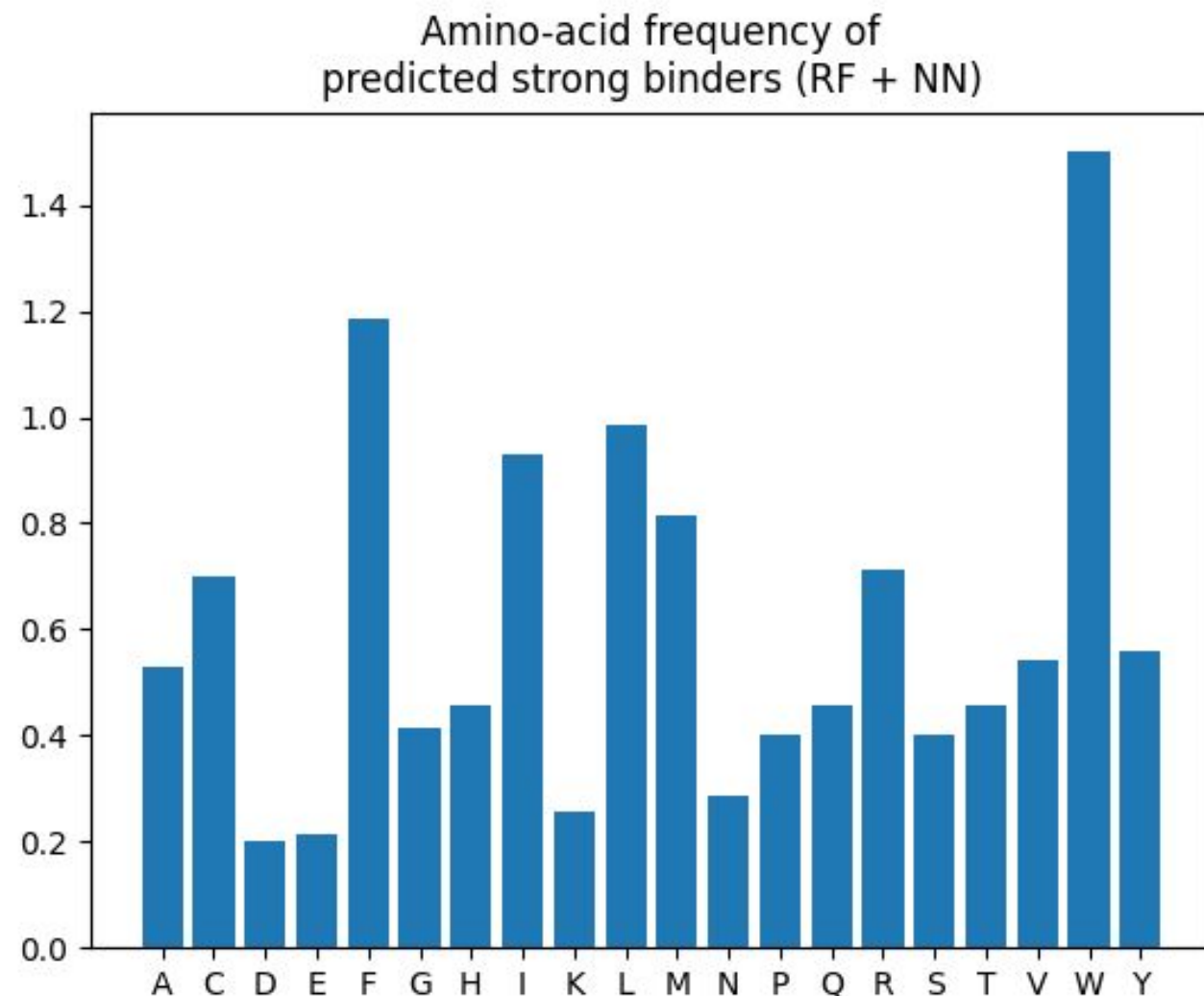
MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

DE NOVO DESIGN - Exploratory Random Search

Results in peptides that contain mostly:

- Aromatic residues
- Cysteine

Peptide	Prediction
WLWSPWMPPMCAN	0.796
LWWFLWWCLNII	0.775
IYWRRACGWQPP	0.769
WKQIWFWQFMRC	0.769
WMYMYHWLLCFS	0.766





DE NOVO DESIGN - Exploratory Random Search

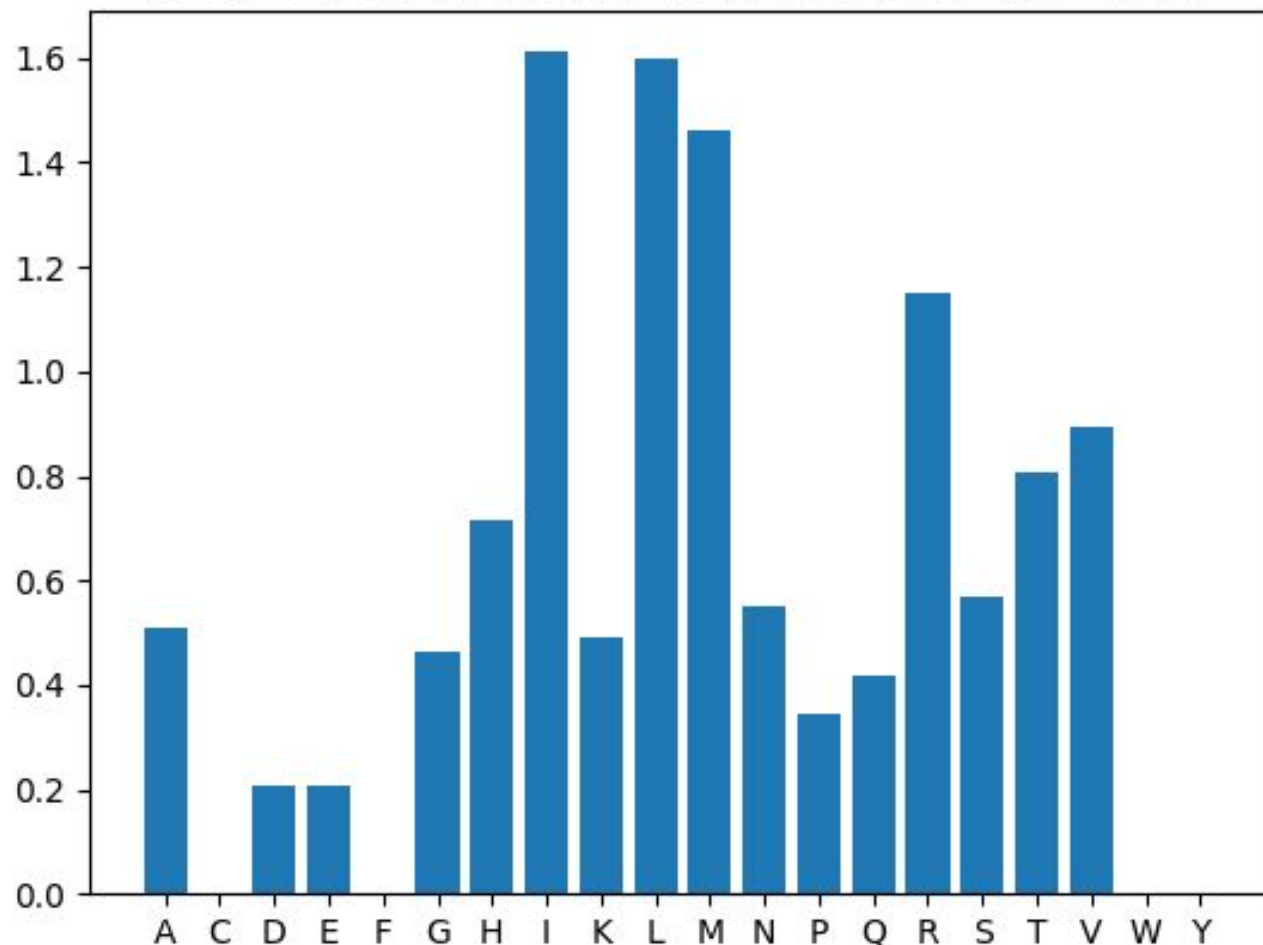
To avoid the non-specificity and the bias*

- Removed aromatic residues
- Removed cysteine

A restricted search resulted in:

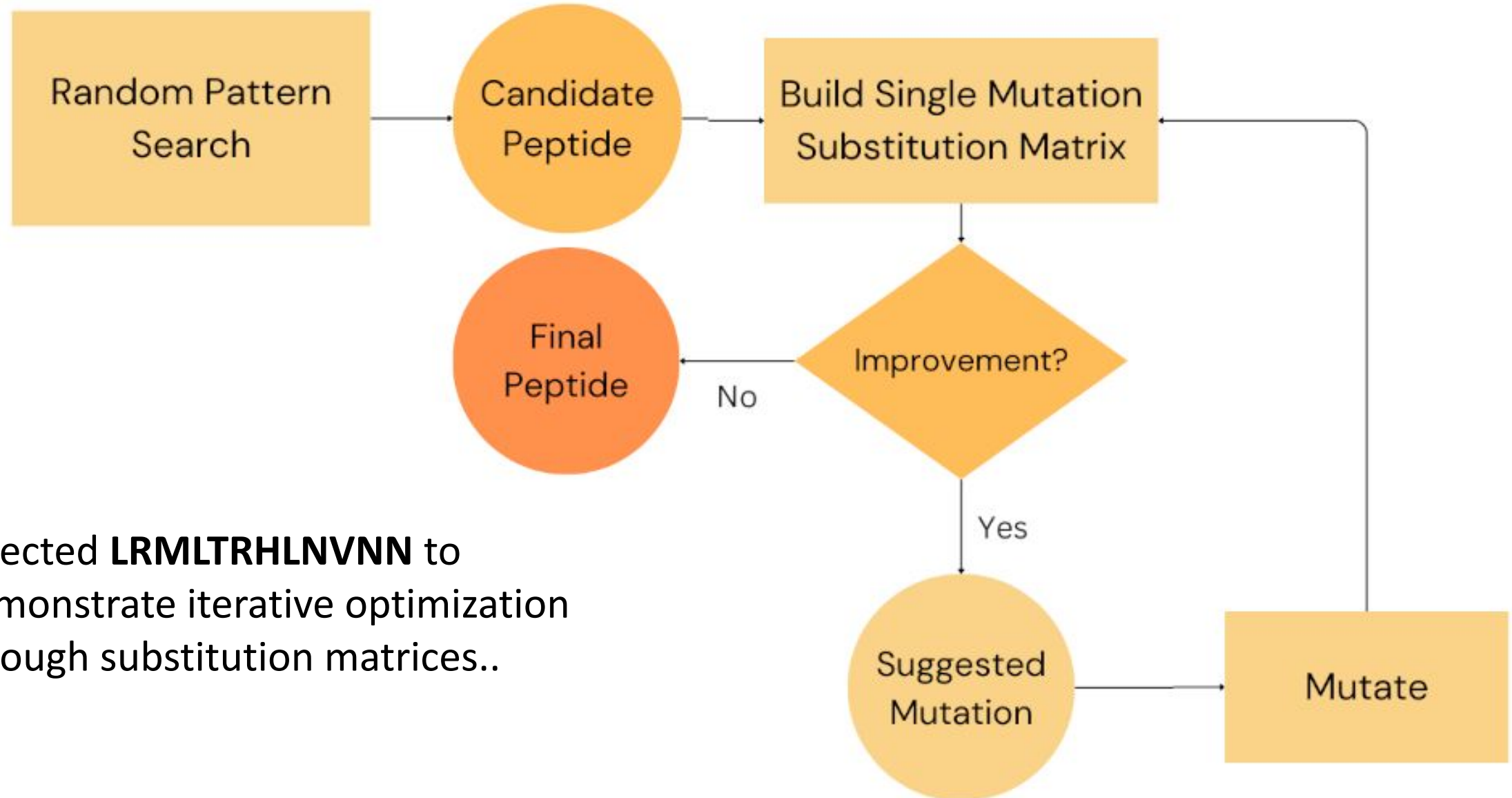
Peptide	Prediction
LRMLTRHLNVNN	0.699
LLVMIGMLKSSQ	0.679
MHLIRHIMGAVM	0.67
LMILGGVMKNVA	0.661
MAIMVRQHEALV	0.661

Amino-acid frequency of predicted strong but non-aromatic binders (RF + NN)





DE NOVO DESIGN - Iterative Optimization



Selected **LRMLTRHLNVNN** to demonstrate iterative optimization through substitution matrices..

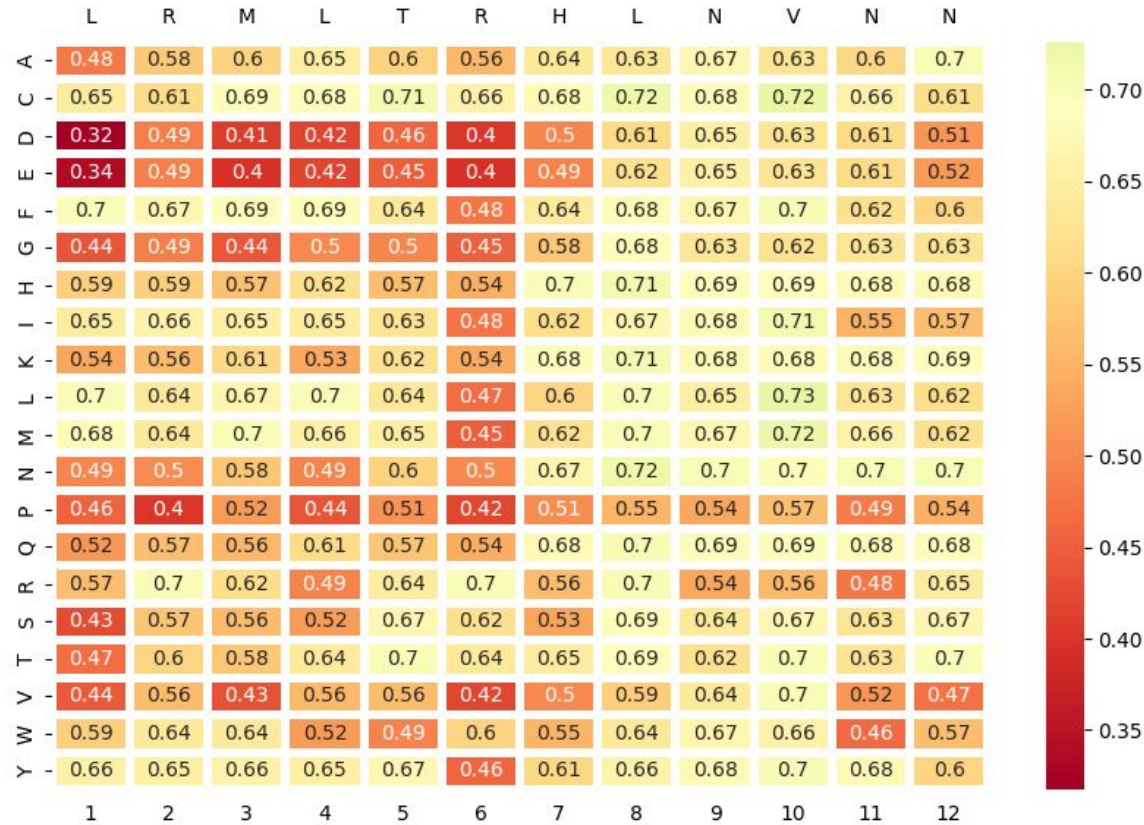


MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

DE NOVO DESIGN

Suggests mutating V10 to L10

Peptide

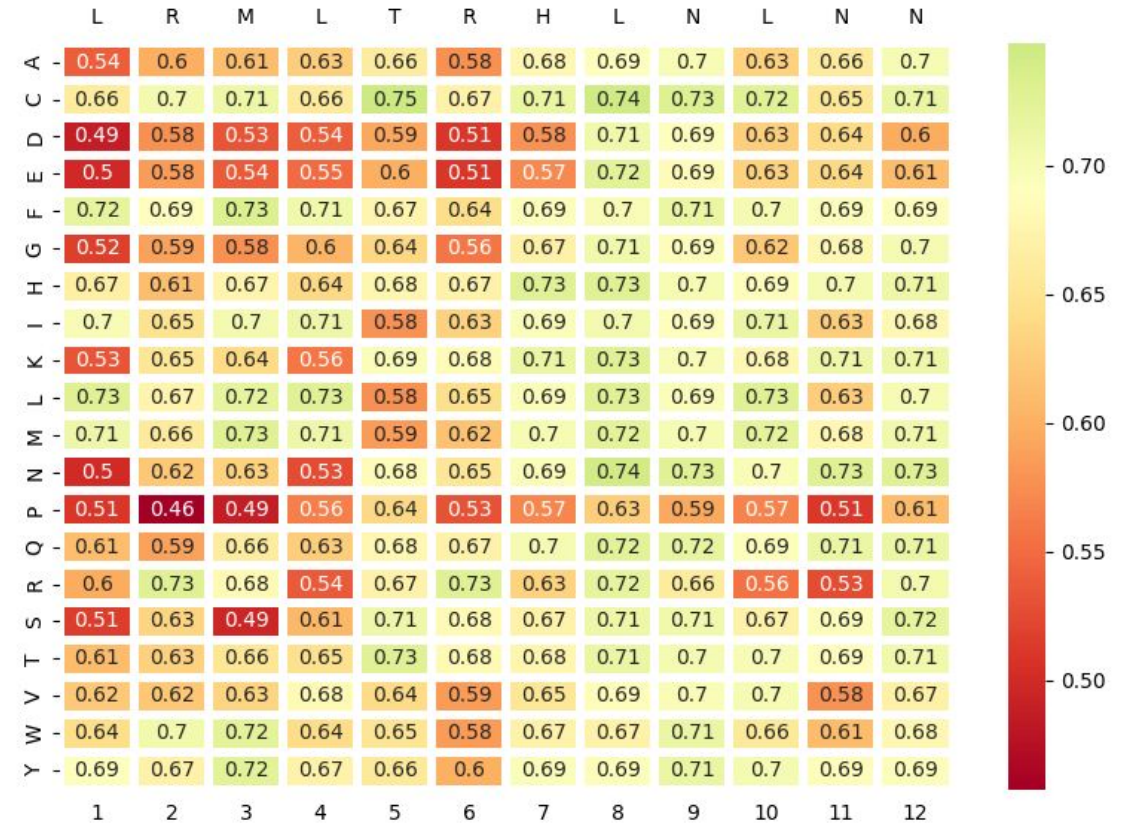


Position

Substitution matrix of the initial candidate

Suggests mutating L8 to N8

Peptide



Position

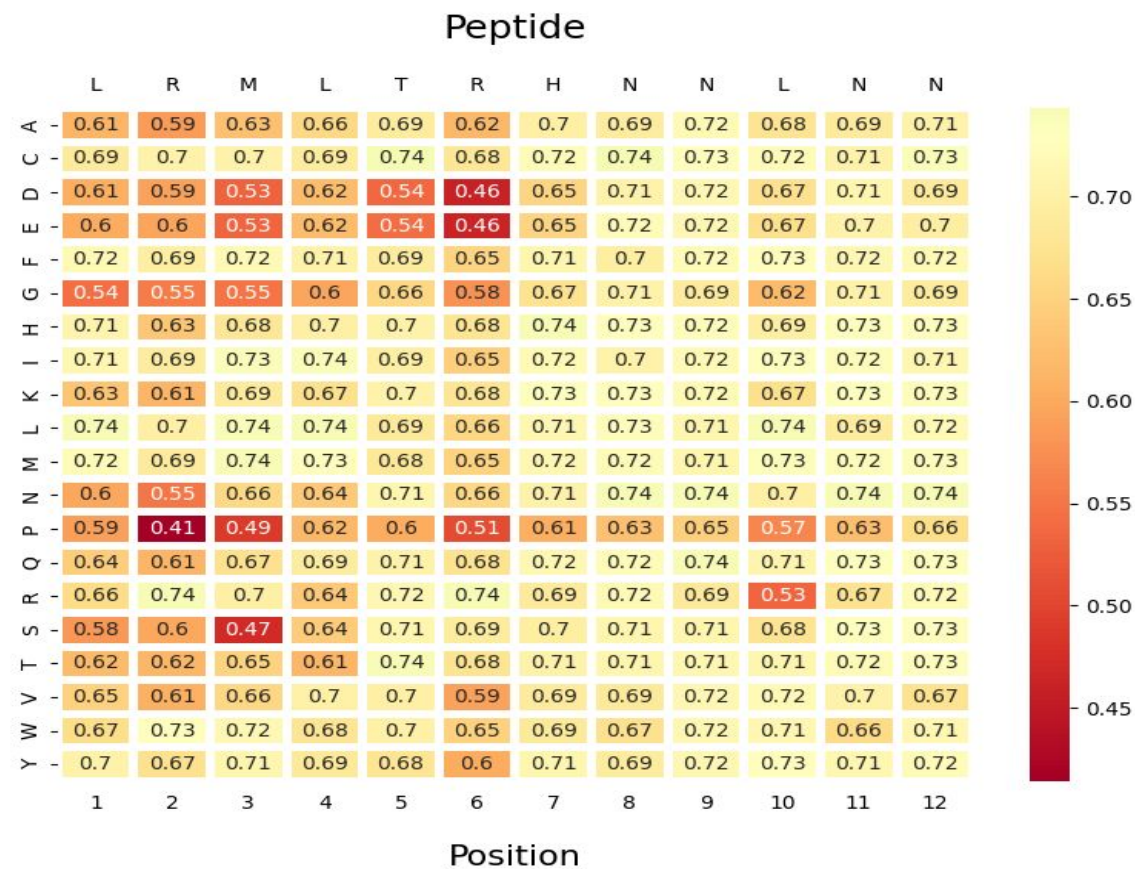
Substitution matrix of the candidate after mutation



MACHINE-LEARNING-ASSISTED DE NOVO DESIGN OF MOLYBDENUM DISULFIDE BINDING PEPTIDES

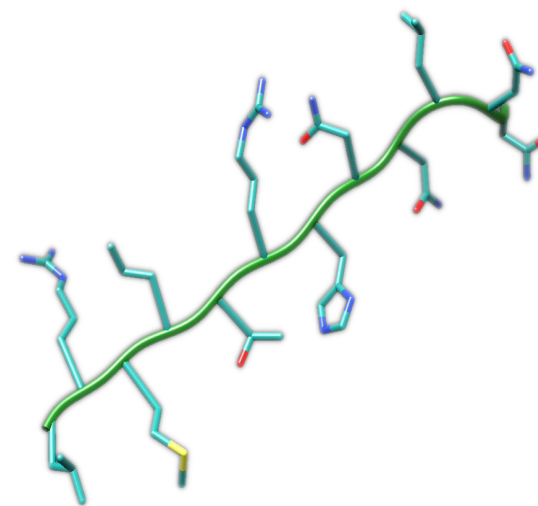
DE NOVO DESIGN

The optimization has ended: No positive gradient

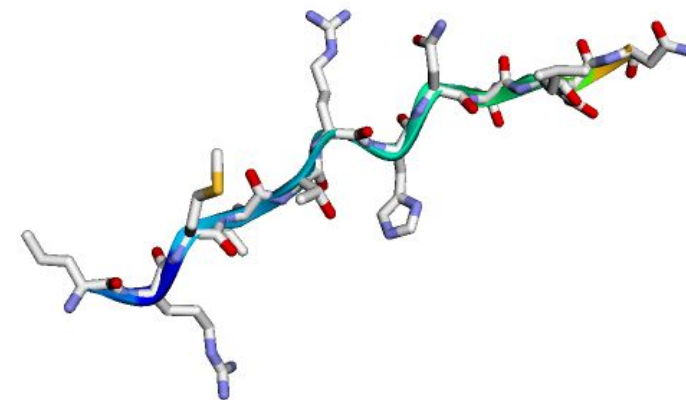


Substitution matrix of the final candidate

Example candidate peptide: LRMLTRHNNLNN
Predicted score: 0.74



3D structure as predicted by
OmegaFold.



3D structure as
predicted by
AlphaFold2.



CONCLUSIONS & FUTURE WORK

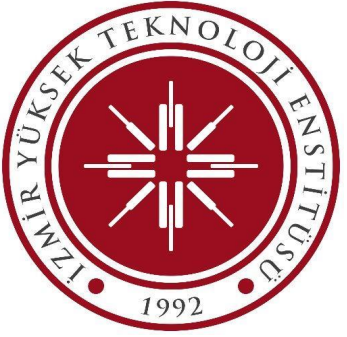
- Explored the data generated by the Deep-directed Evolution experiment.
- Built and compared machine learning models that predicts MoS₂ binding affinity through the data.
- Built a peptide design workflow ready for any high-throughput phage display experiment.
- Deep-directed evolution approach proves to be revolutionary in peptide/protein design.

Future work:

- Explore more efficient peptide search. E.g. genetic algorithm.
- Validate with other public phage display data.
- Search peptides with certain patterns, motifs, and similarities with natural proteins



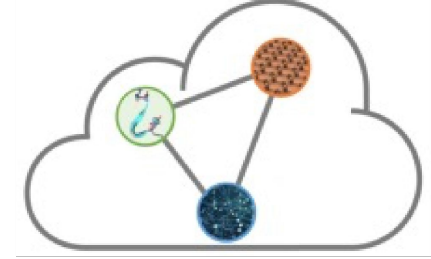
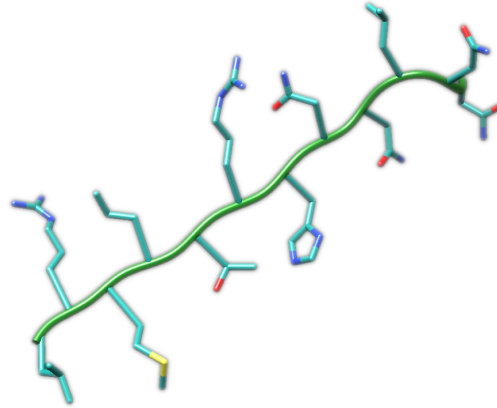
ACKNOWLEDGEMENTS



BIOTECHNOLOGY

Team Members

İlker Kan- MSc. Student in Biotech
Nursevim Çelik- MSc. Student in BioE
Gizem Çulha - MSc. Student MatE
Alara Coşkun- UG Student BioE
Kerem Haznedar - UG Student MatE
Eda Yurdamil Cica - UG Student BioE
Yağmur Çeşmeci- UG Student BioE
Bengisu Bağcı- UG Student BioE
Şevval Çığ- UG Student BioE



YUCESoy RESEARCH GROUP

Funding Agencies



Horizon 2020_MSCA-IF; REDEEM 10102960053



TÜBİTAK
2210-C

IYTE BAP; 2021IYTE-1-0114; 2022IYTE-2-0050