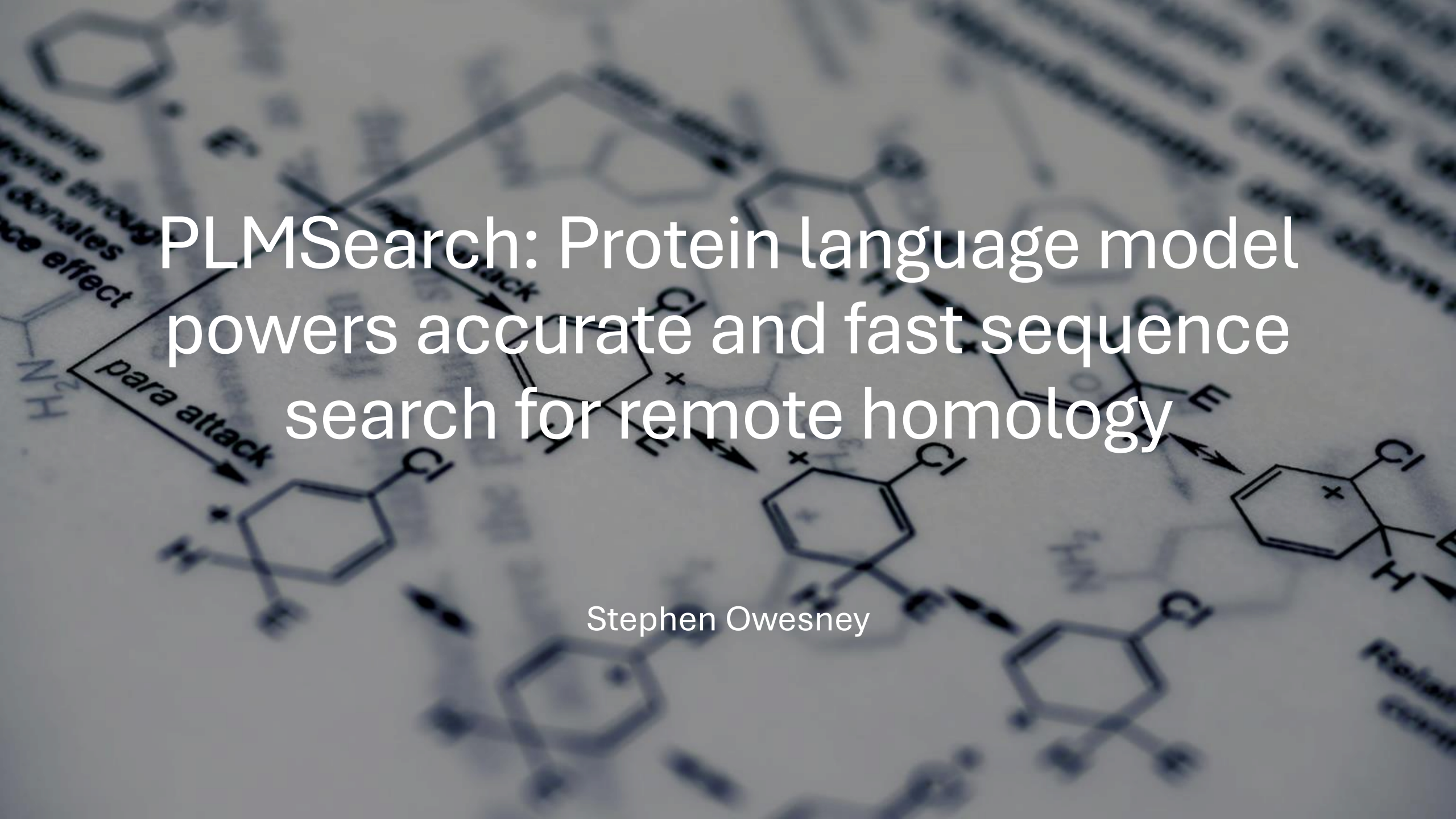
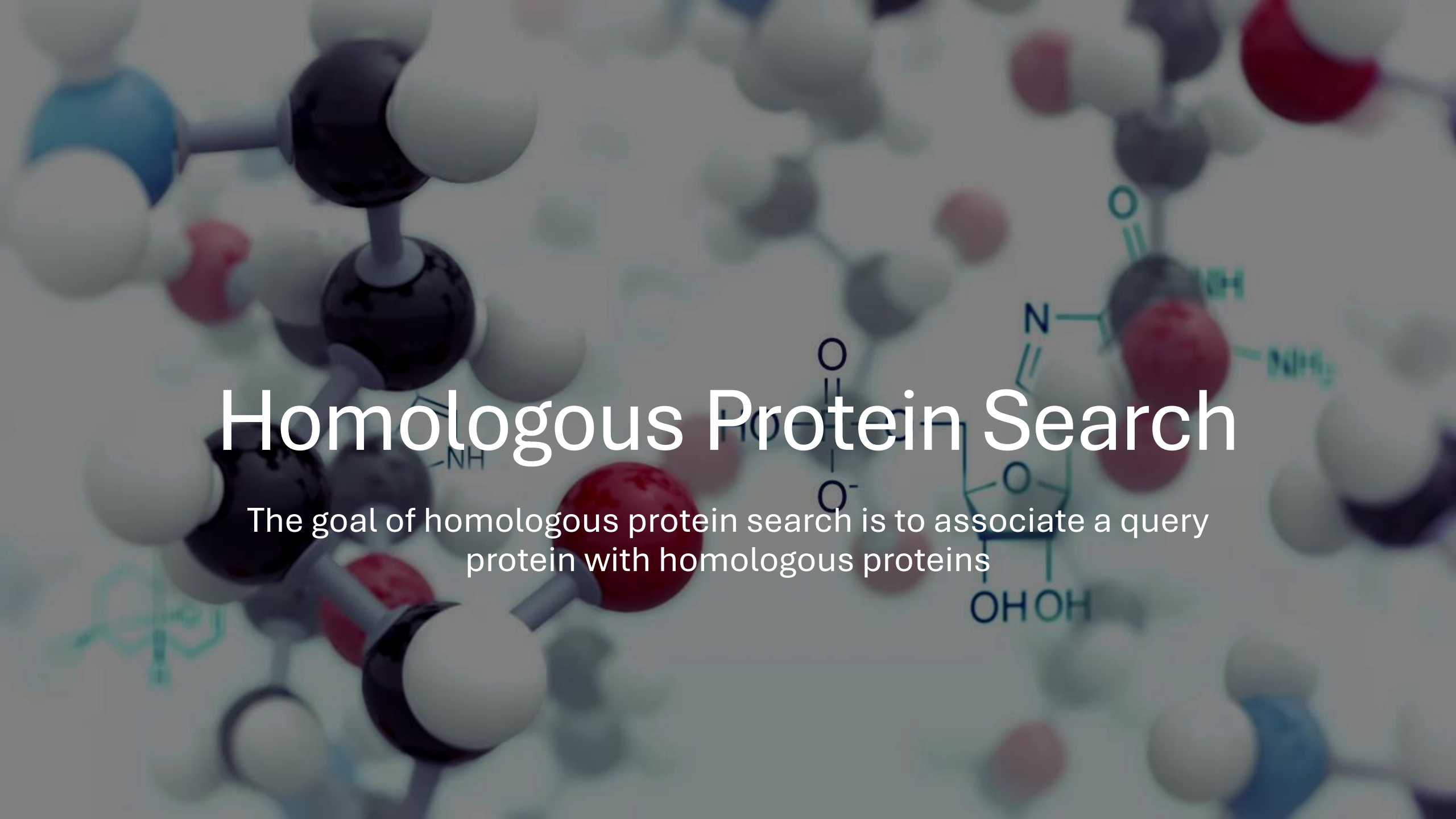


PLMSearch: Protein language model powers accurate and fast sequence search for remote homology

Stephen Owesney

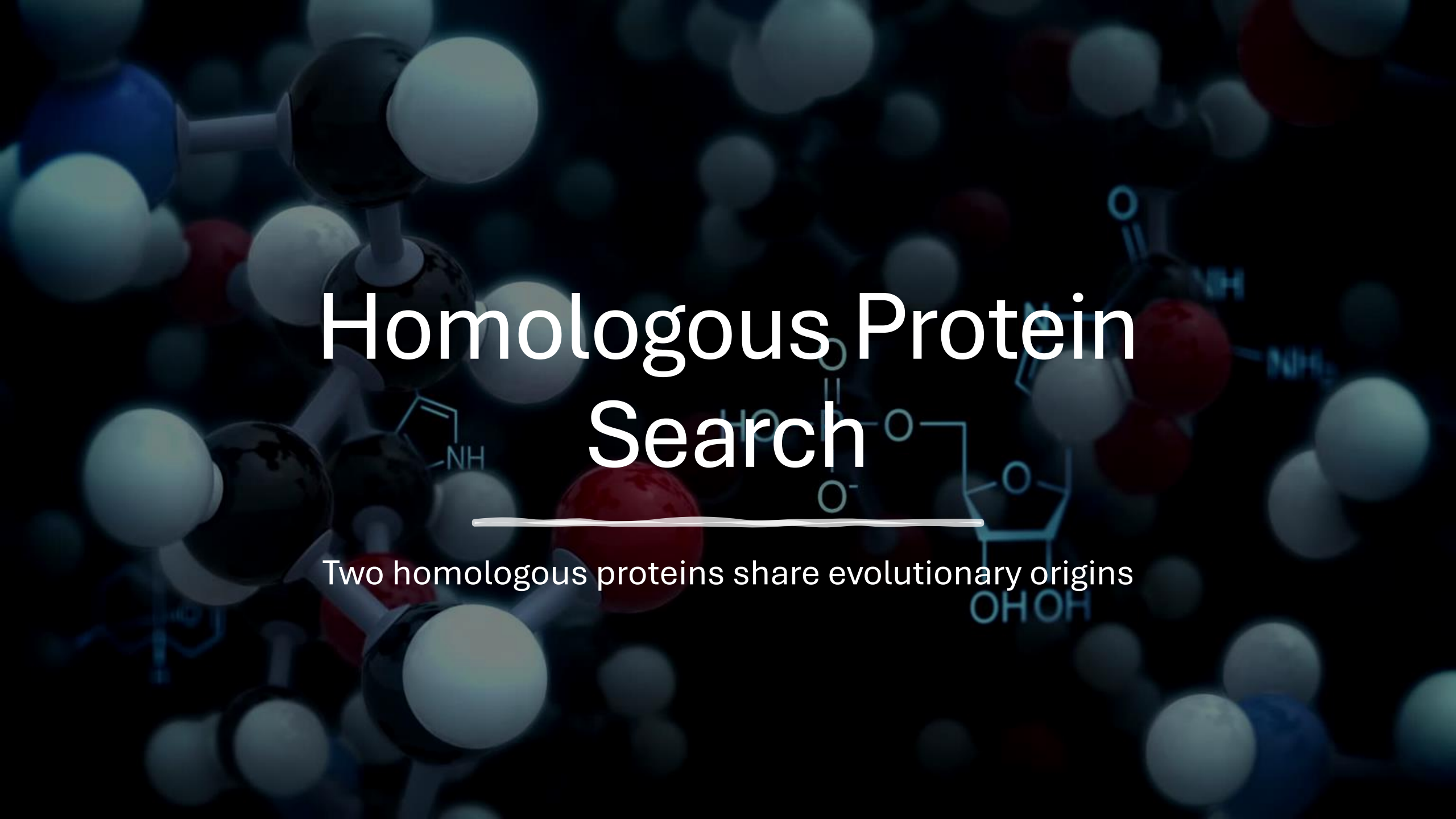
Overview

- Introduction to Homologous Protein Search
- Background Knowledge / Related Works
- Introduction to PLMSearch
- Methodology
- Results / Experiments
- Future Work
- Questions / Discussion



Homologous Protein Search

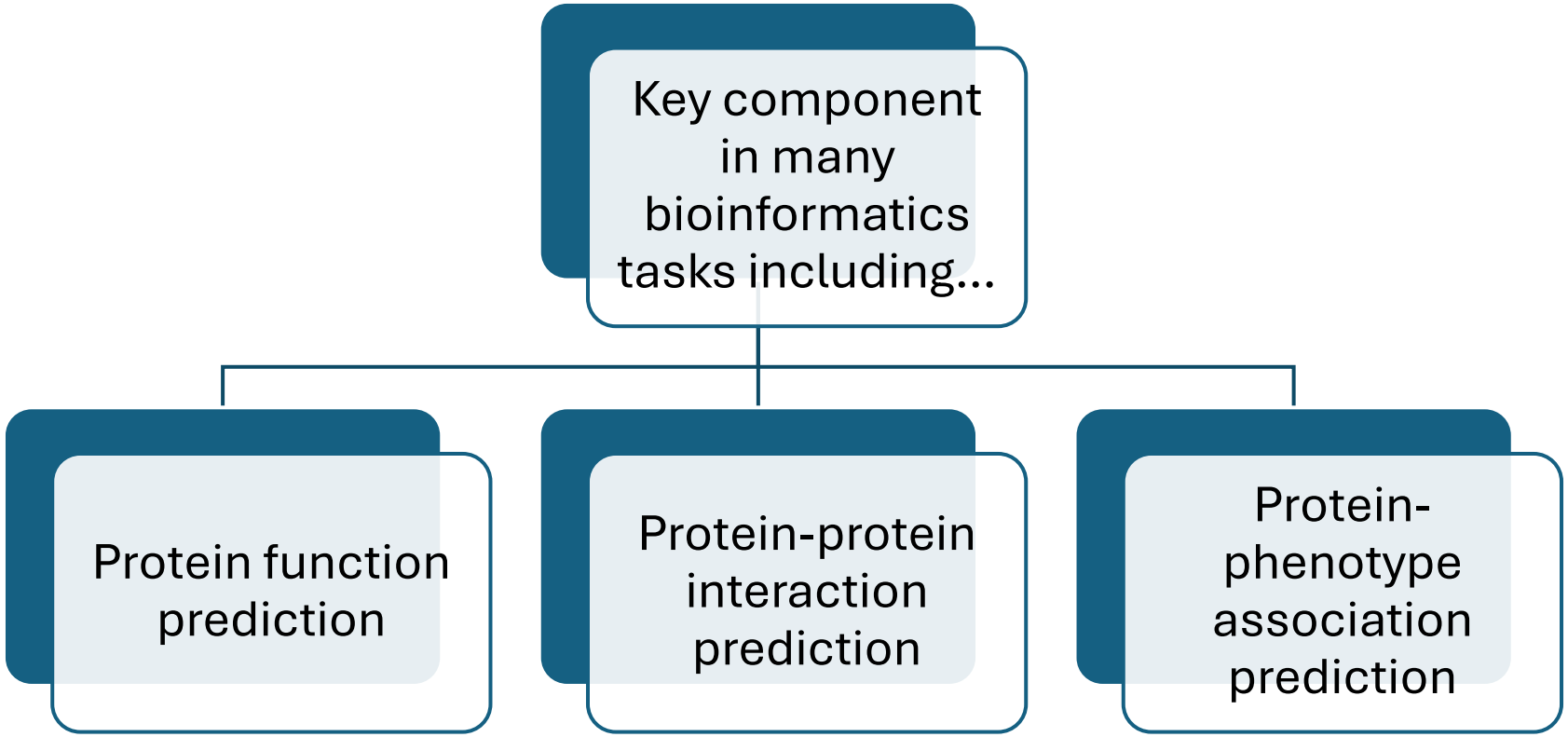
The goal of homologous protein search is to associate a query protein with homologous proteins



Homologous Protein Search

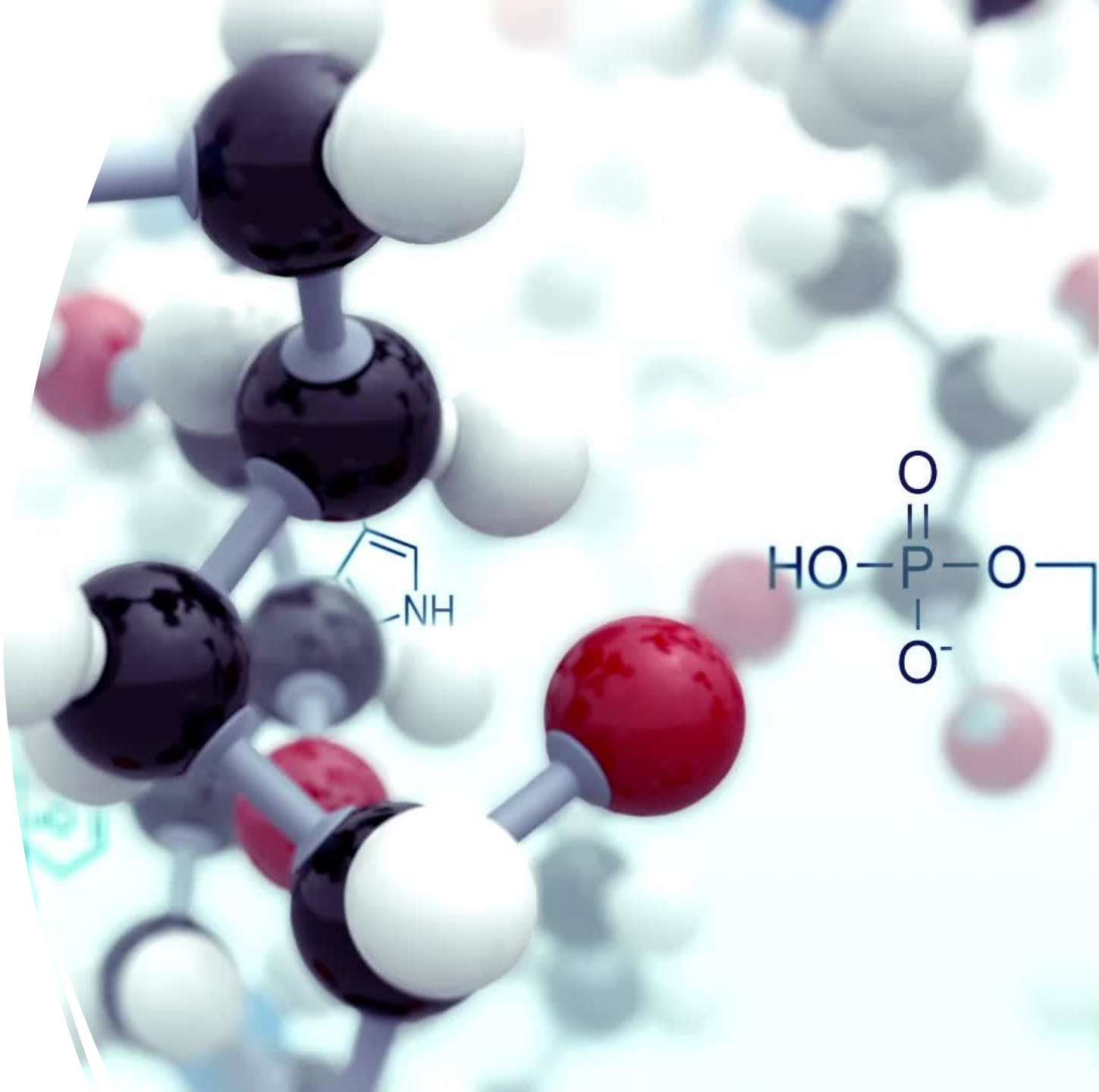
Two homologous proteins share evolutionary origins

Homologous Protein Search



Key Challenges for Homologous Protein Search

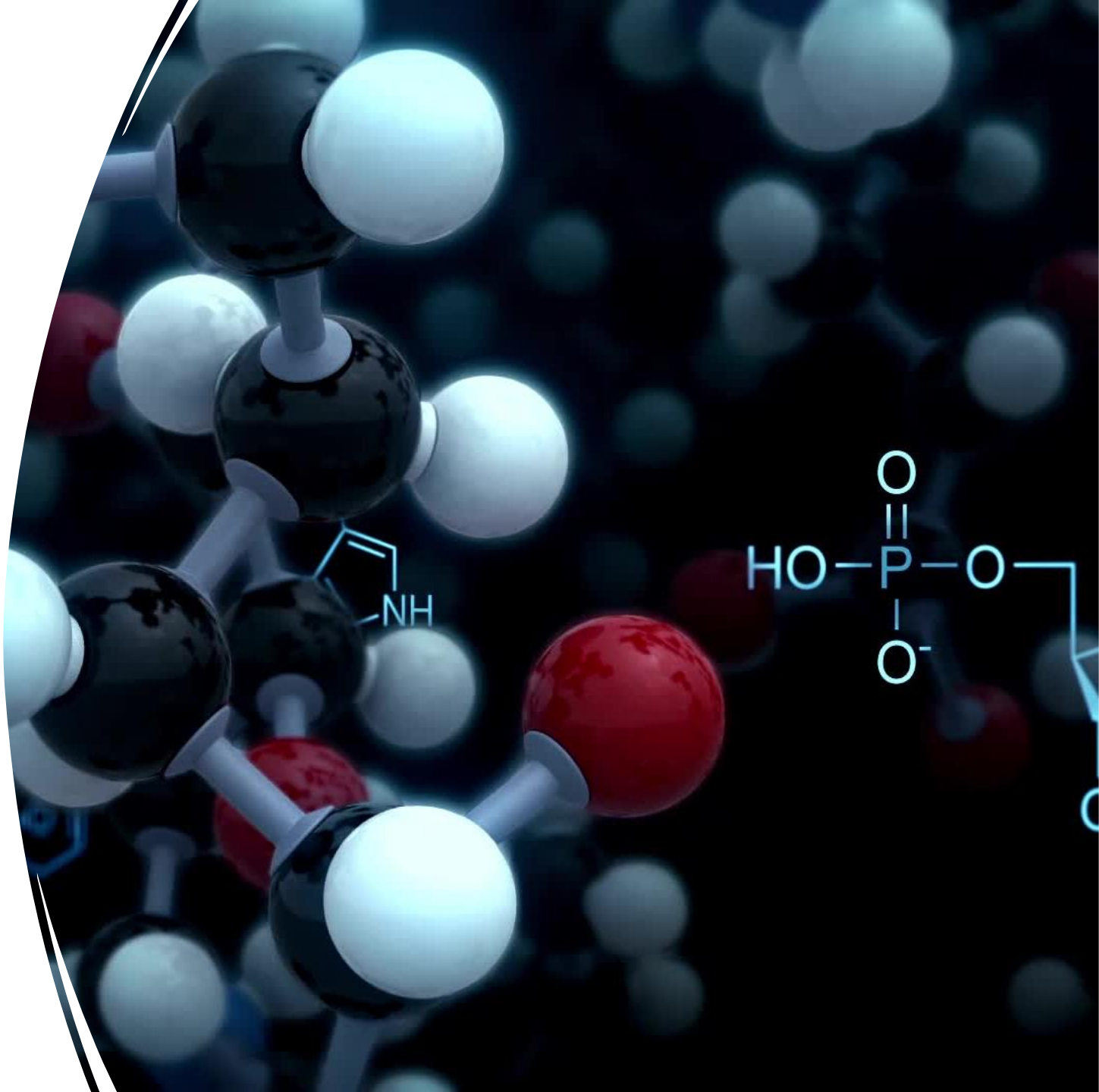
1) Computational Complexity
2) Noisy Datasets
3) Low Sequence Similarity



Homologous Protein Search

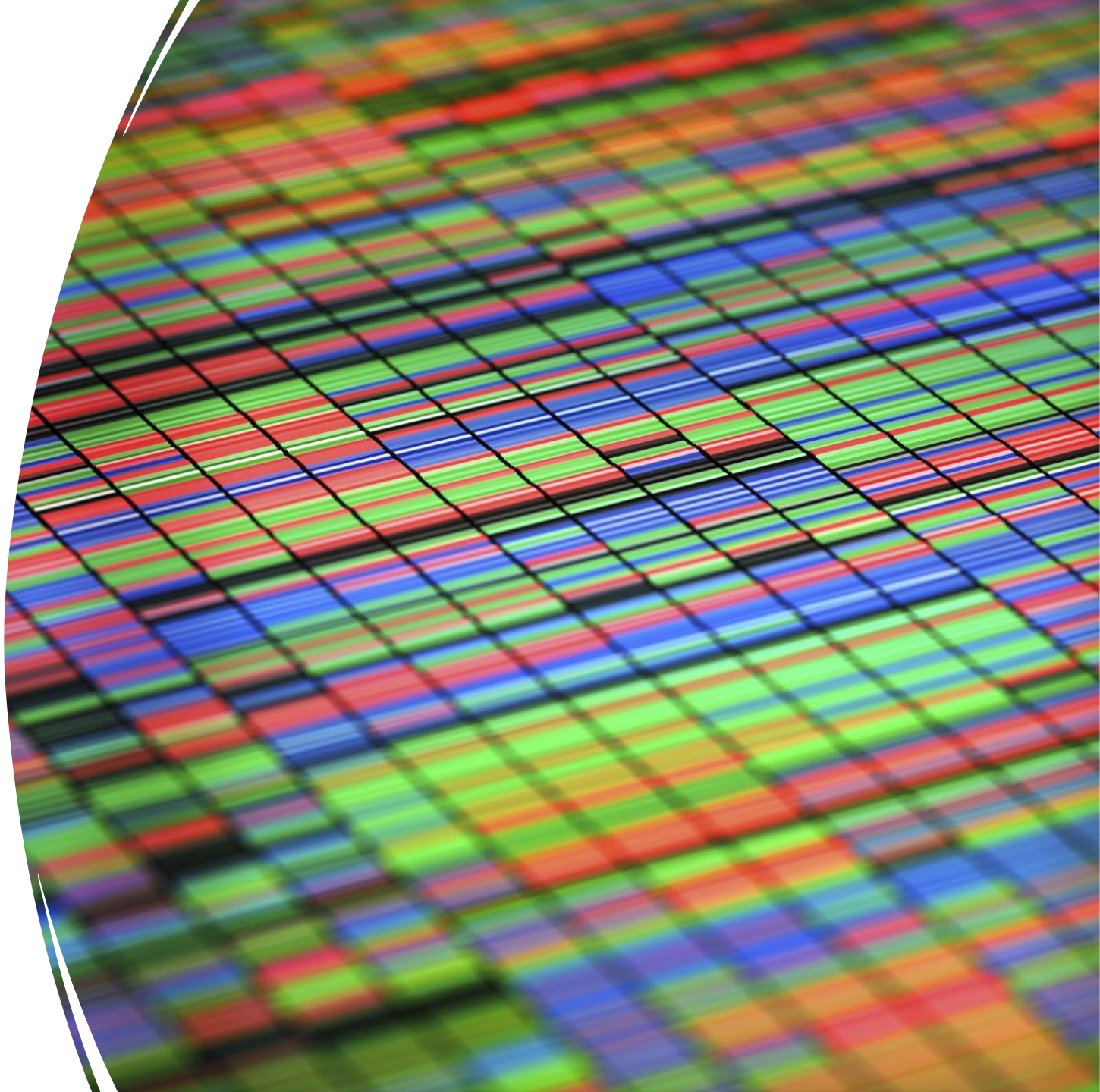
Homologous protein search can be roughly divided into two approaches or components

1) sequence search
2) structure search



Sequence Search Approach

Goal: Compare amino acid sequence of query to targets in database



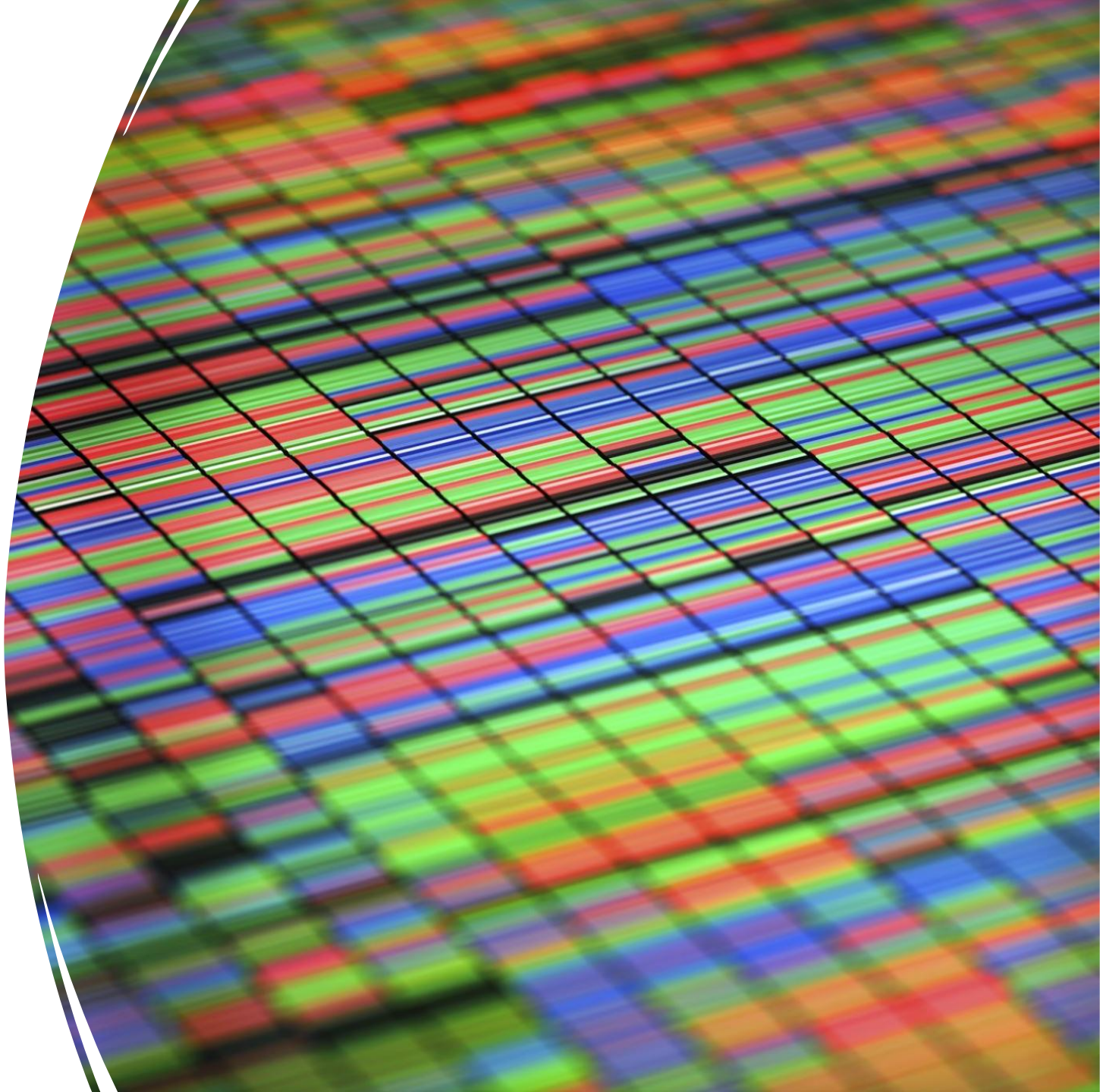
Sequence Search Approach

Strengths

1) Low cost
2) Data abundance
3) Well established

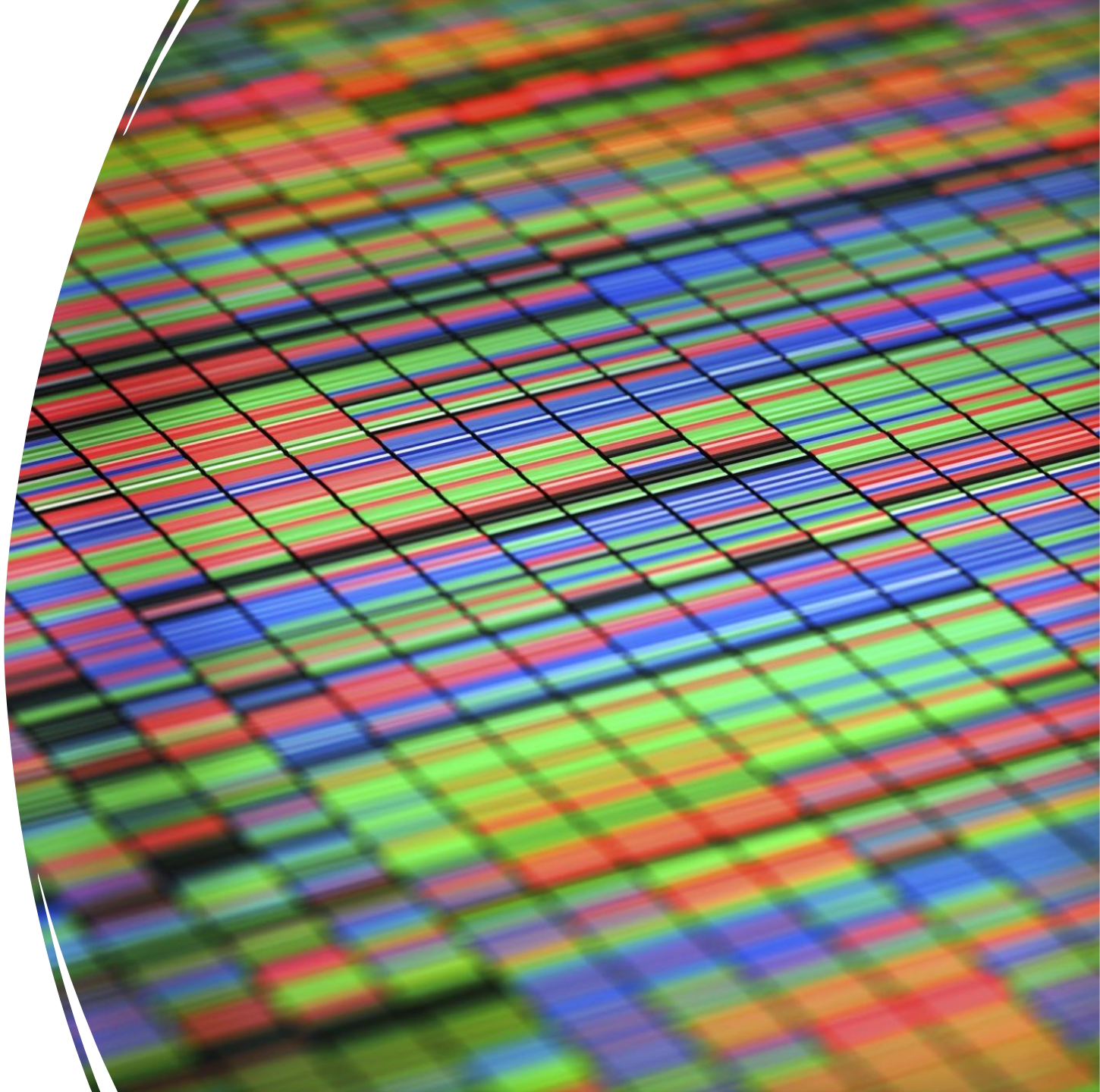
Challenges

1) Sequence divergence
2) Poor remote performance



Sequence Search Approach

- MMseqs2
- BLASTp
- Diamond



Structure Search Approach

- **Goal:** Perform the comparison of the query and target proteins leveraging 3D structure of the proteins

Structure Search Approach

Strengths

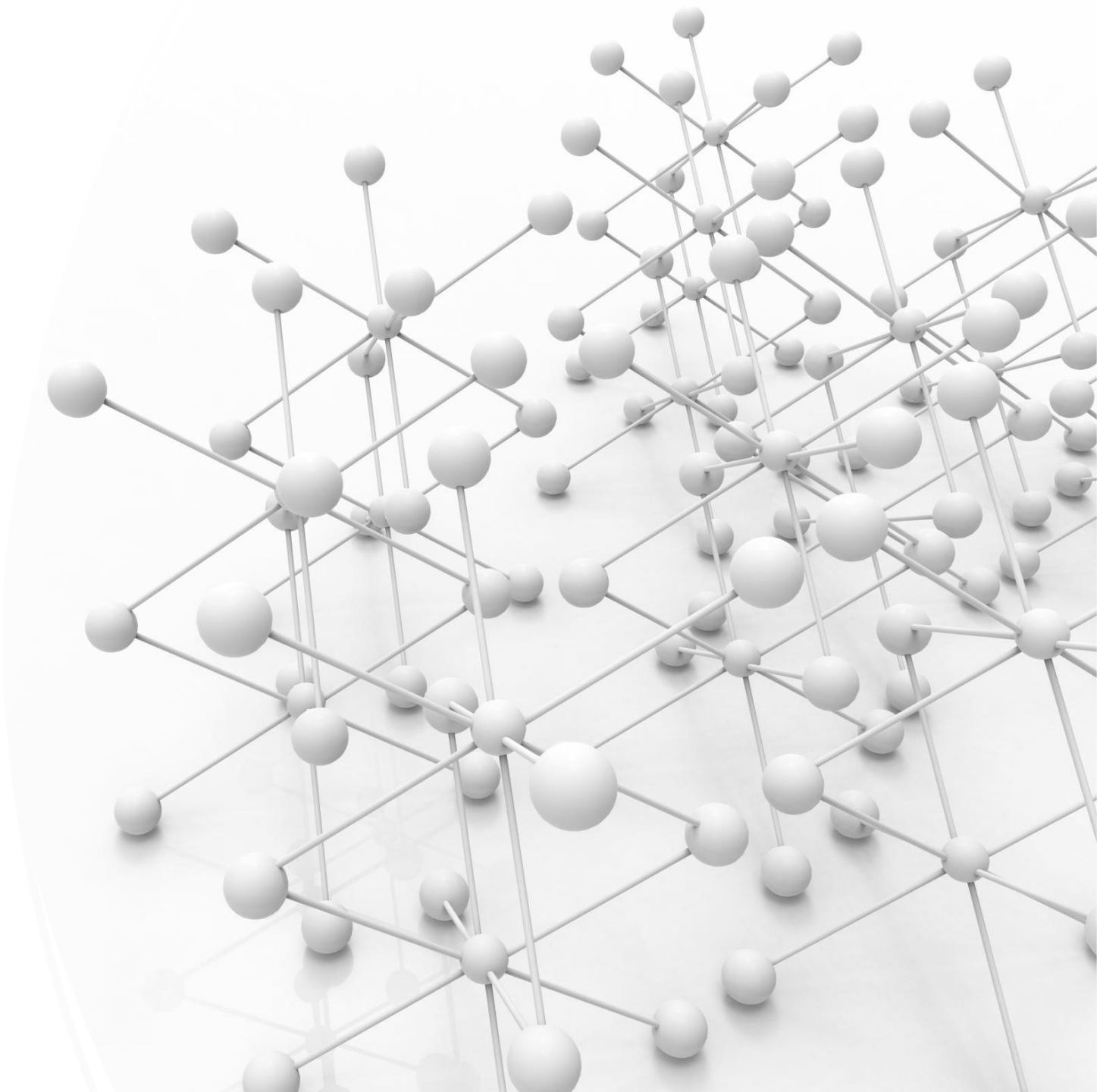
1) Better remote performance

Challenges

1) High cost
2) Data scarcity

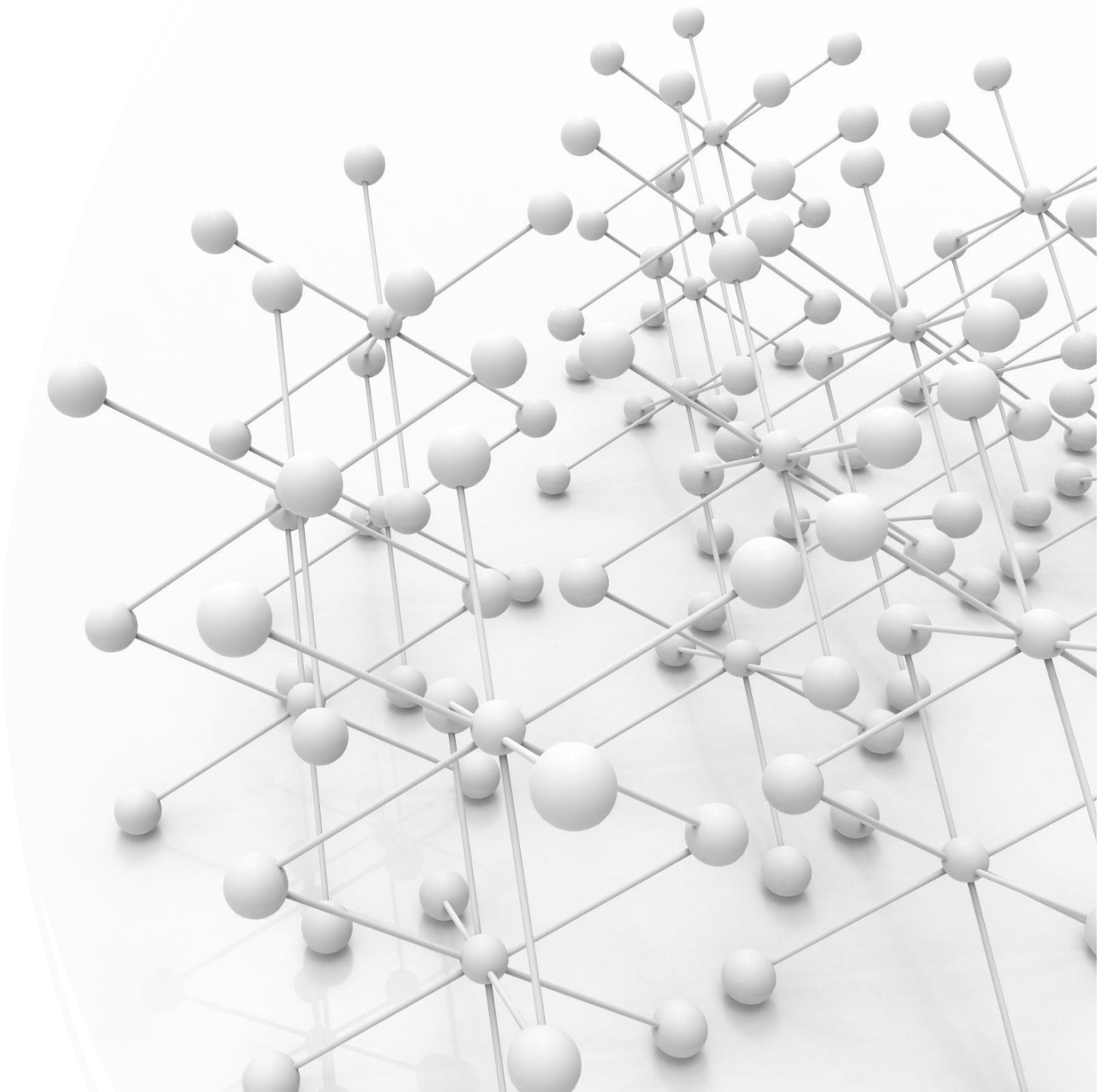
Structure Search Approach

Protein structure search methods can be further divided into 3 main approaches



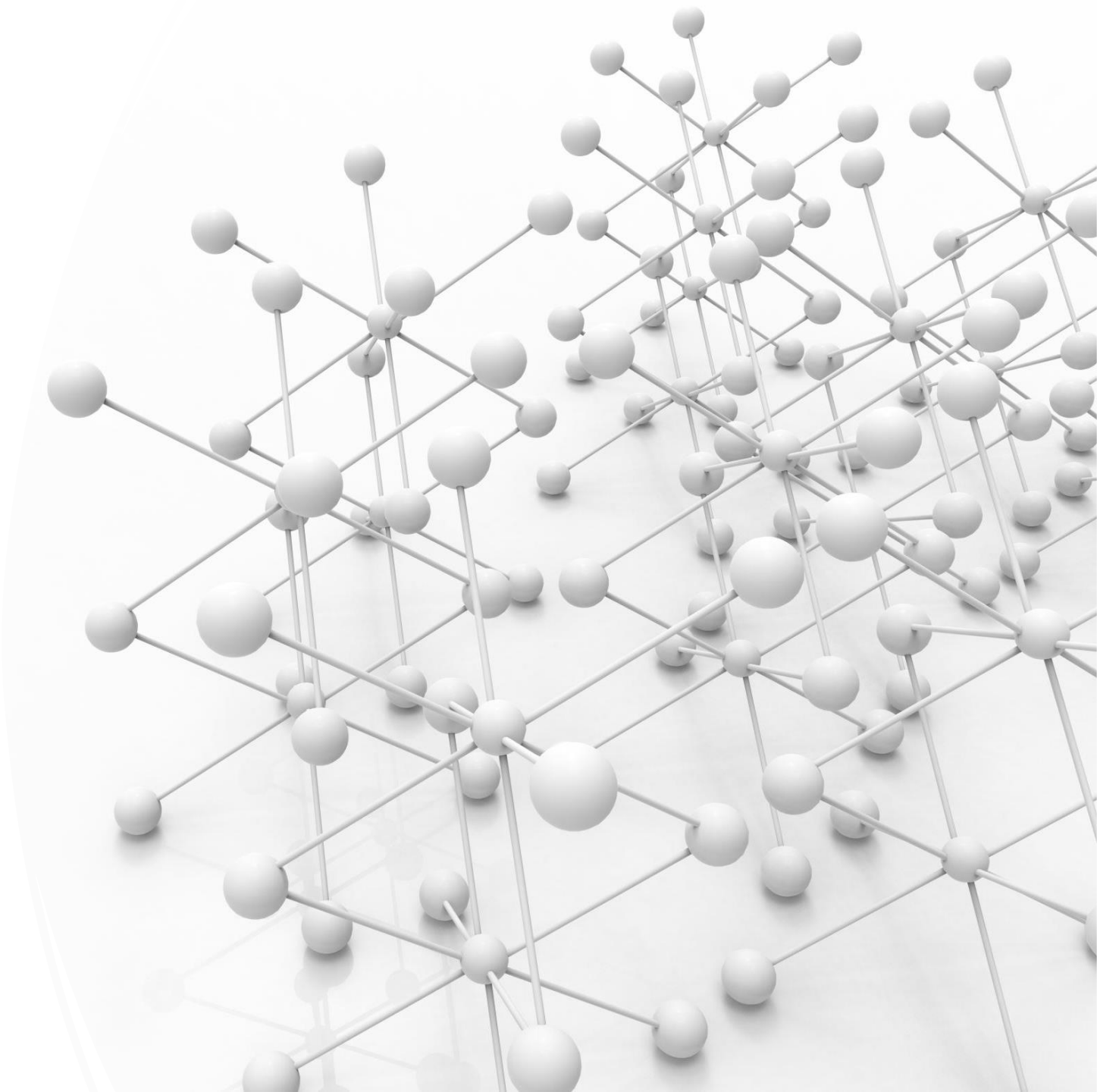
Contact/Distance Map-Based

- Map_align
- EigenTHREADER
- DiscoVER



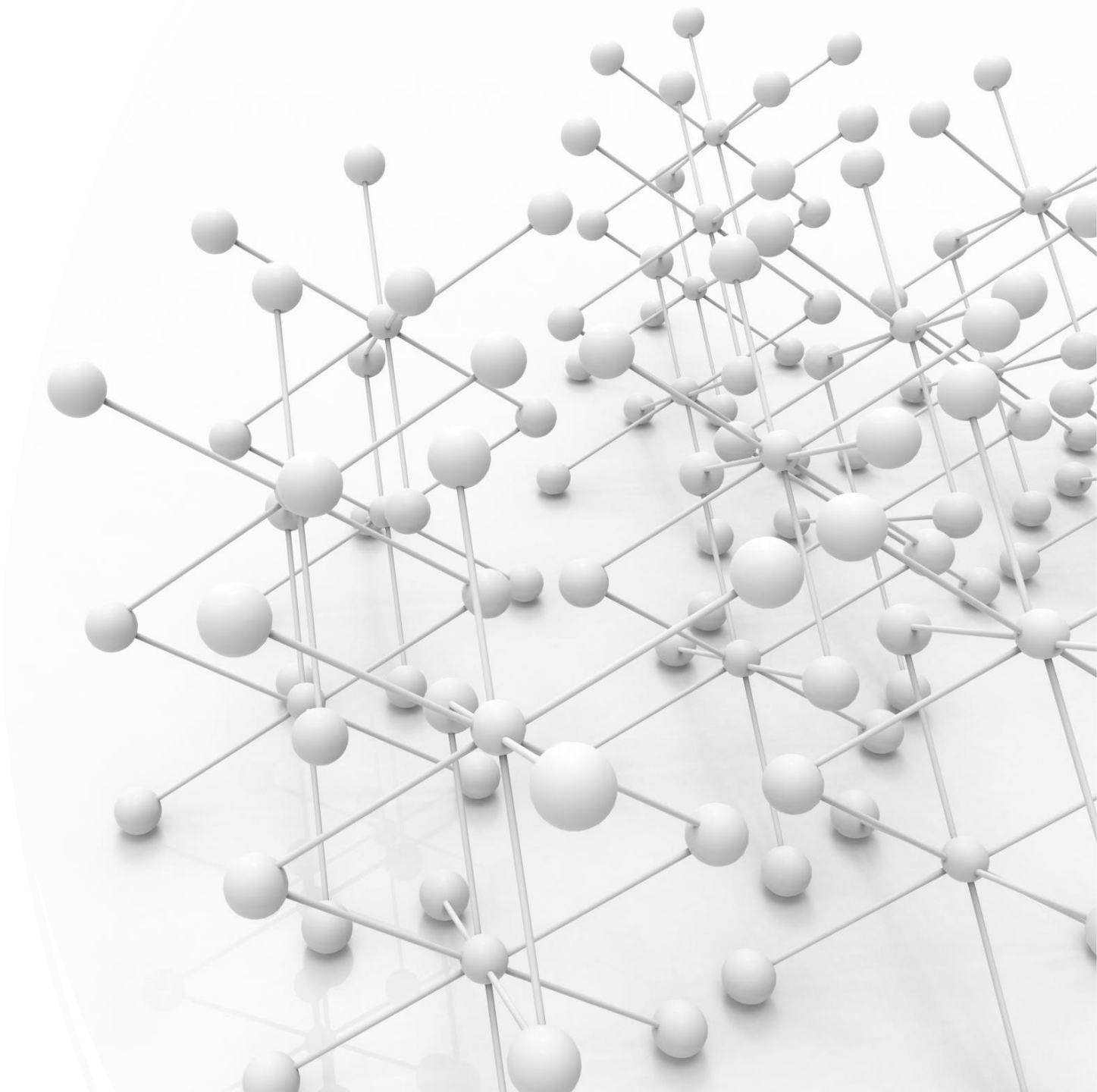
Structural Alphabet-Based

- 3D-BLAST-SW
- CLE-SW
- Foldseek
- Foldseek-TM



Structural Alignment-Based

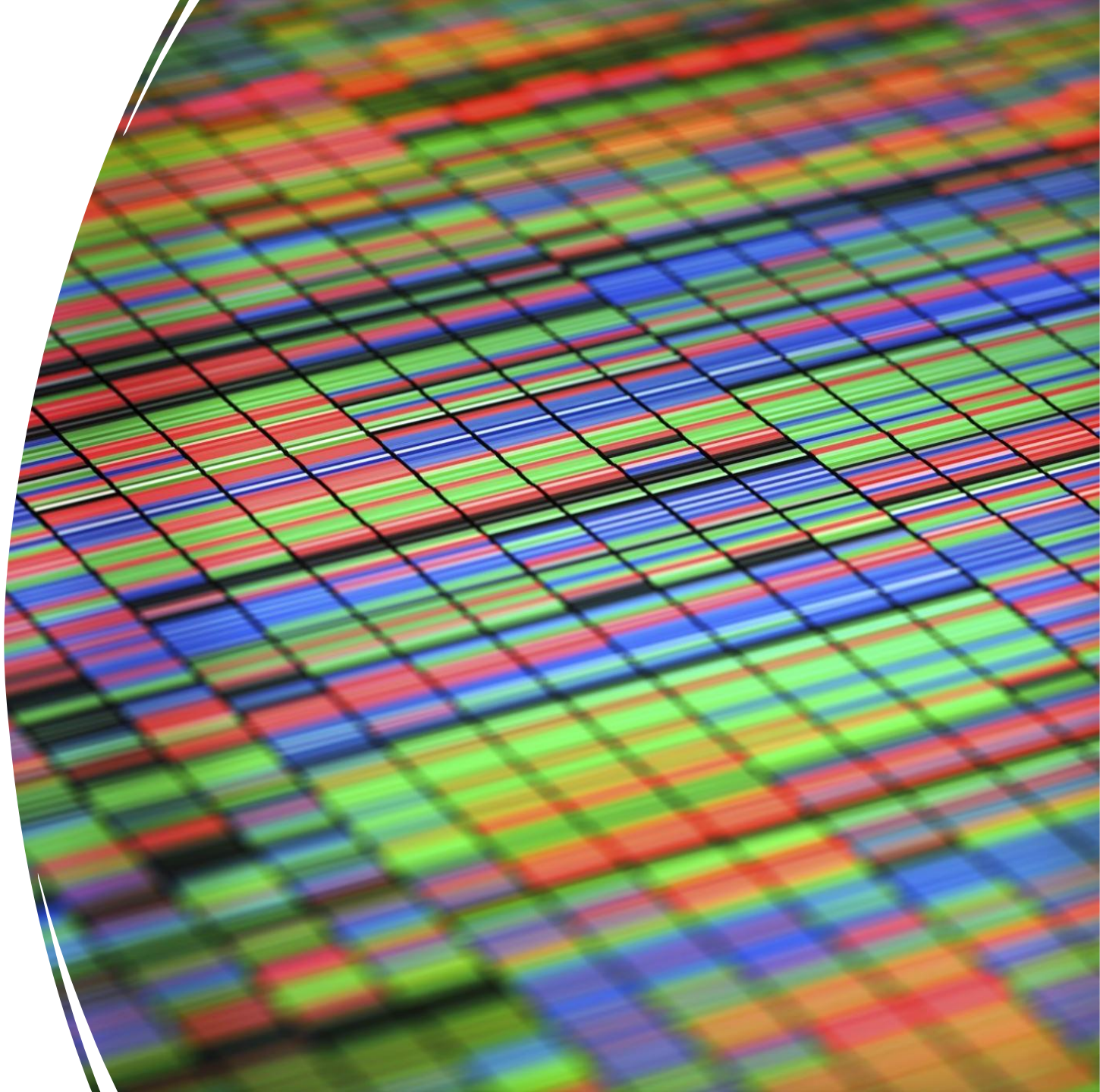
- CE
- Dali
- TM-align

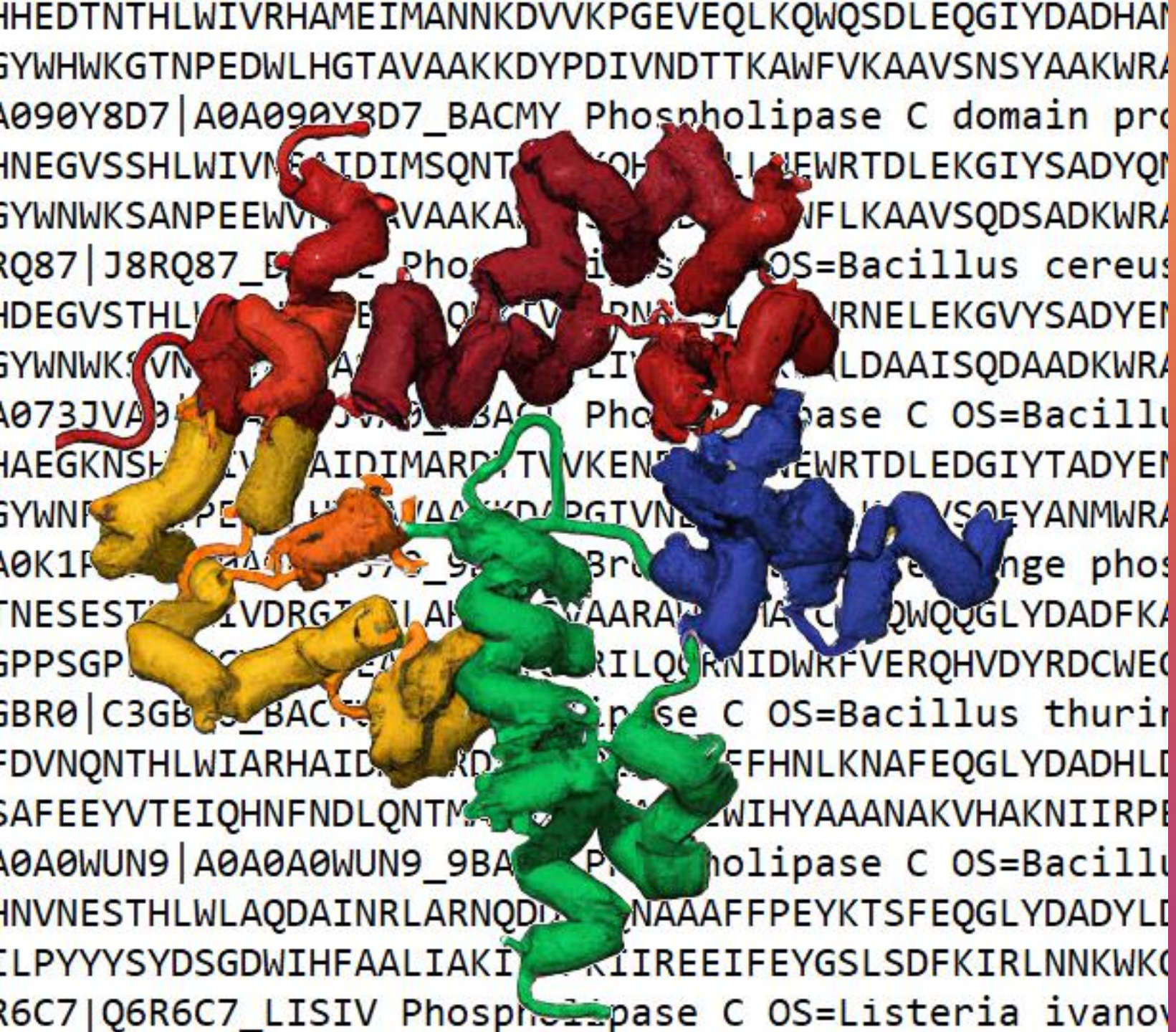


An Old Fix: HMM

Hidden Markov Models (HMMs)

- HMMER
- HHsearch
- HHblits

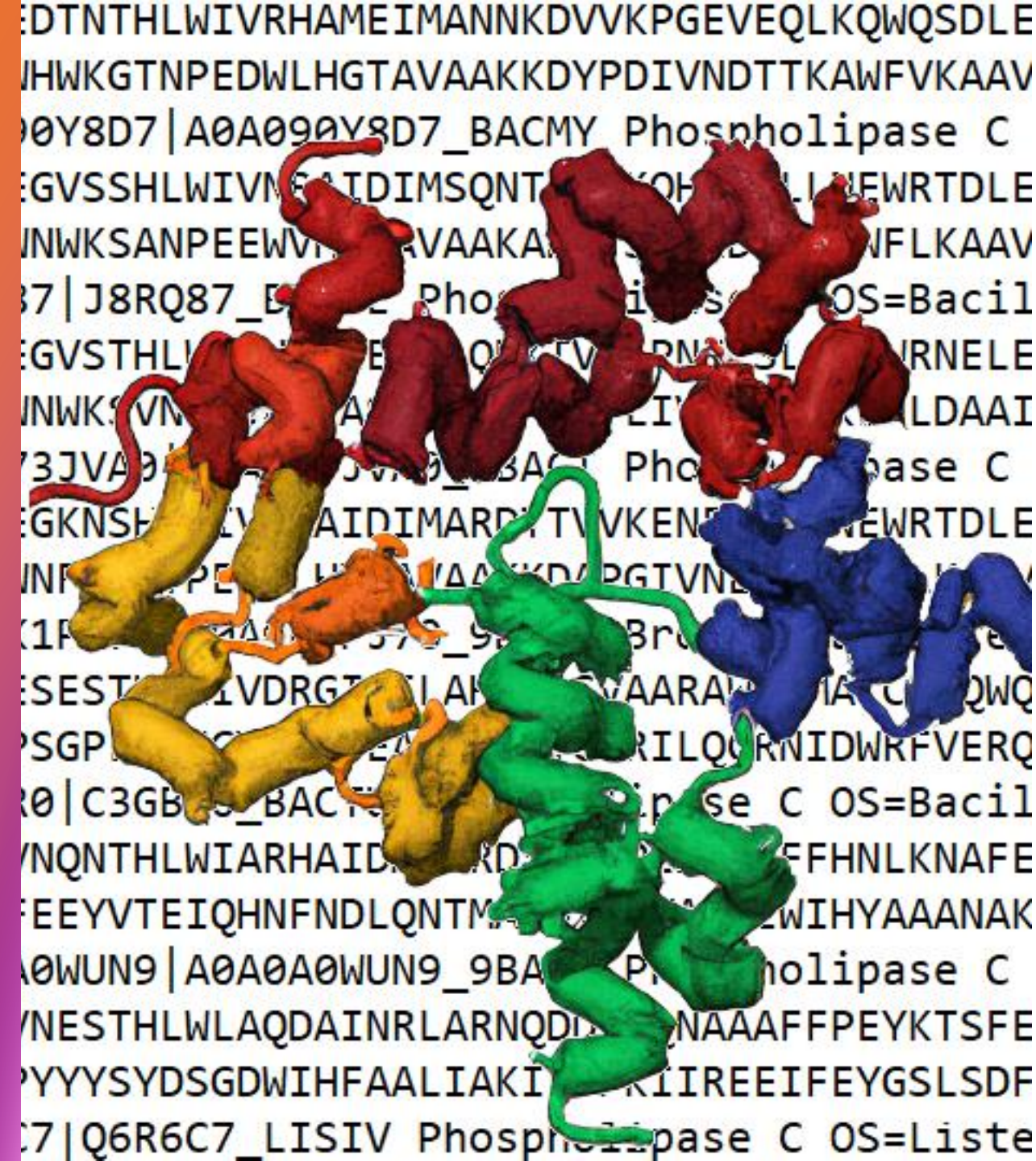




The New Wave: Protein Language Models

(PLM)

Goal: Learn contextual representations of proteins by treating sequences as language



Protein Language Models

- ESMs
- ProtTrans
- ProtENN
- CATHe
- DEDAL
- DeepBLAST
- pLM-BLAST

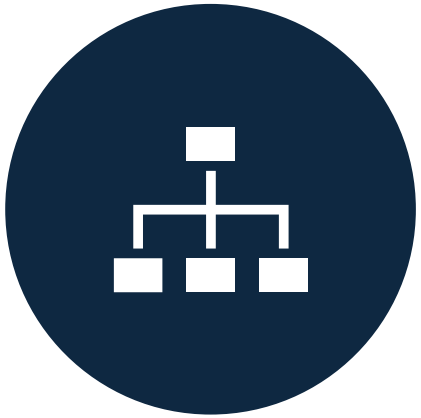
PLMSearch

PLMSearch is designed to improve the **sensitivity** of sequence-based protein searches while maintaining their **efficiency** and **universality**.

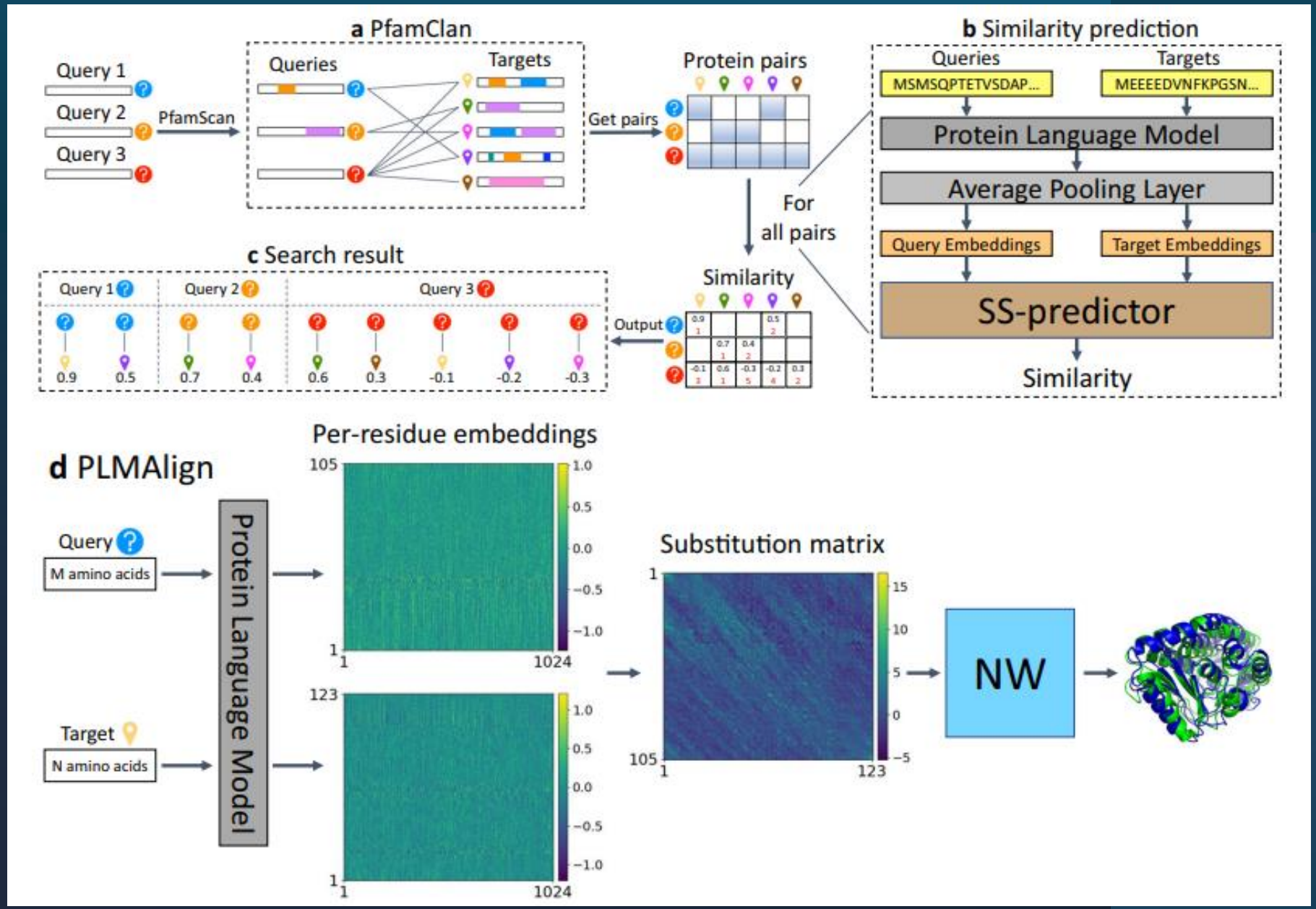
PLMSearch High Level

1) PFAMCLAN FILTERING

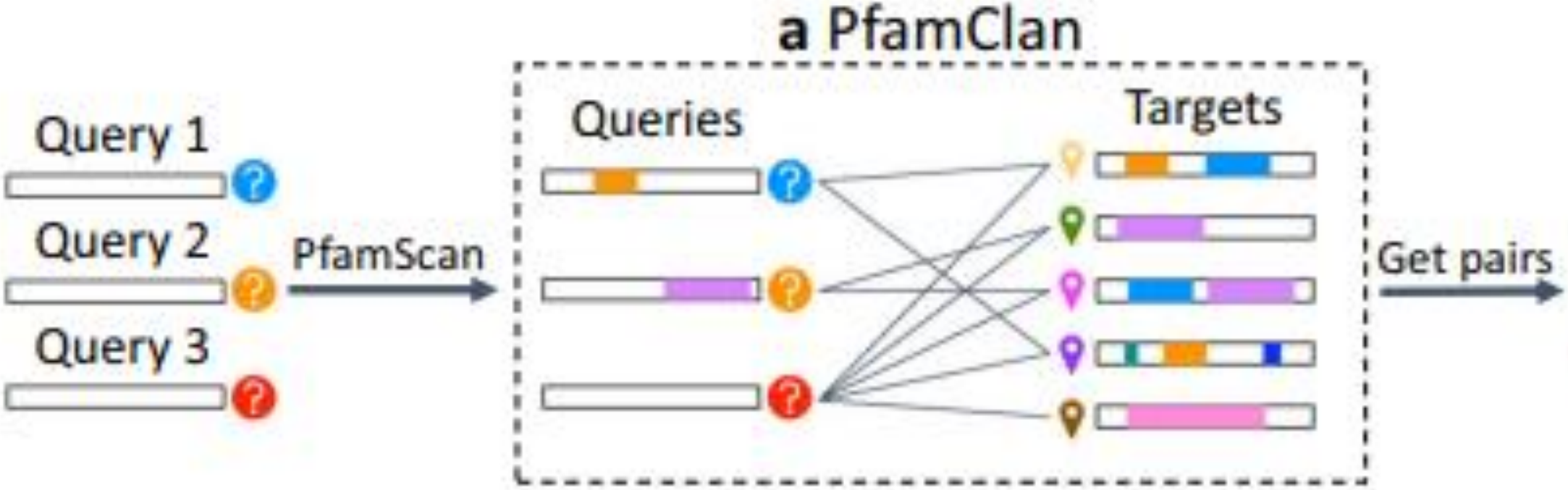
2) SS-PREDICTOR



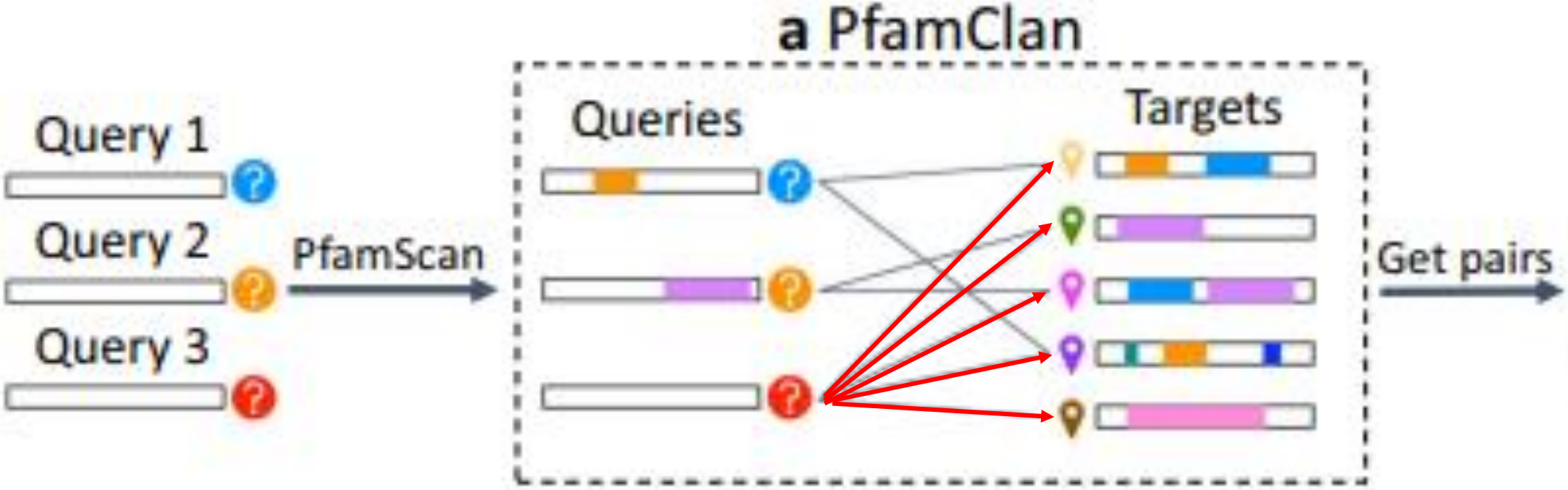
3) RANKING / PLMALIGN


a PfamClan
Query 1
Query 2
Query 3
PfamScan
Queries
Targets
Get pairs
Protein pairs
b Similarity prediction
Queries
Targets
MSMSQPTETVSDAP...
MEEEEDVNFKPGSN...
Protein Language Model
Average Pooling Layer
Query Embeddings
Target Embeddings
SS-predictor
Similarity
For
all pairs
c Search result
Query 1
Query 2
Query 3
0.9
0.5
0.7
0.4
0.6
0.3
-0.1
-0.2
-0.3
Output
Similarity
d PLMAlign
Query
M amino acids
Target
N amino acids
Protein Language Model
Per-residue embeddings
105
1
1024
123
Substitution matrix
NW

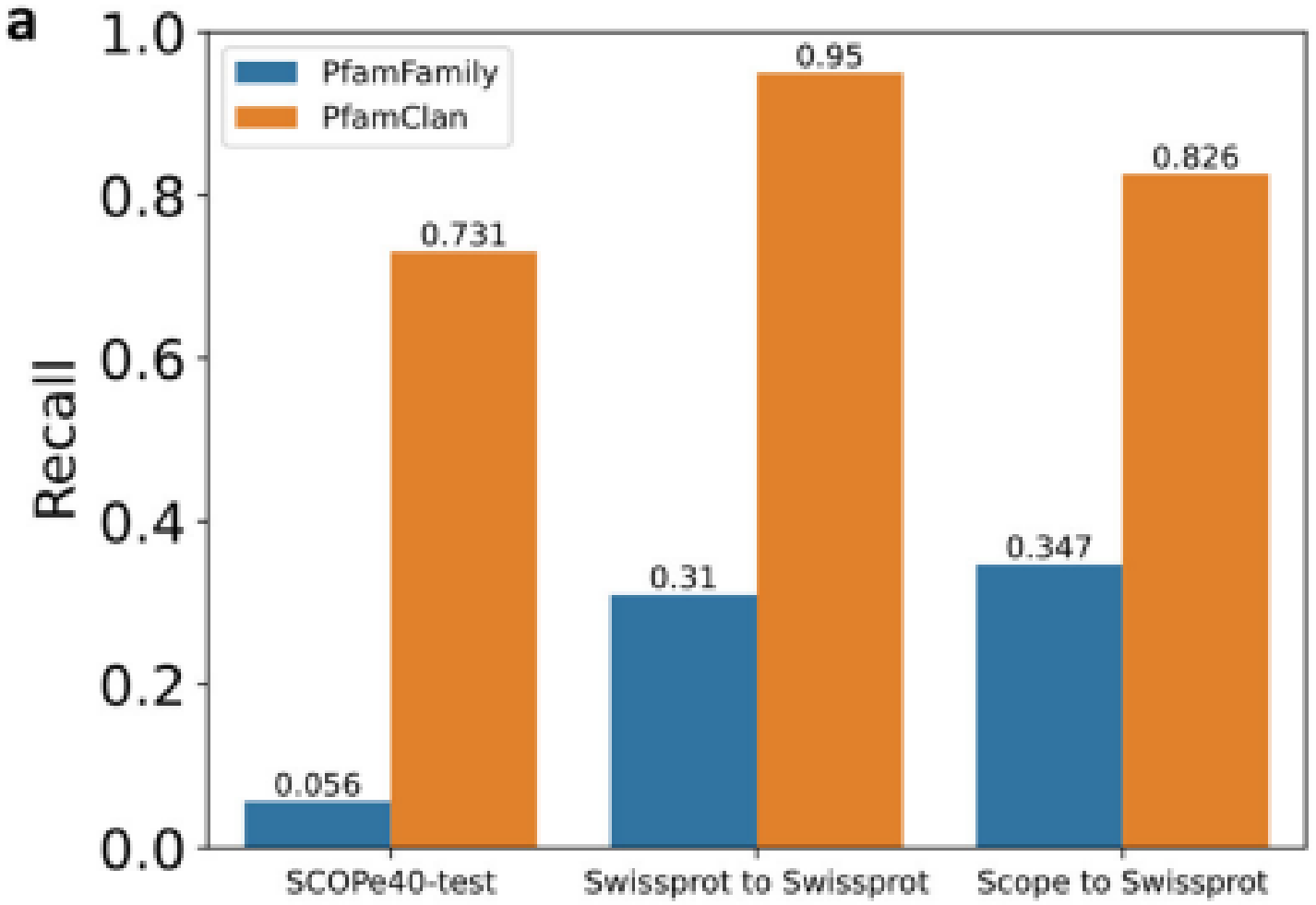
PfamClan

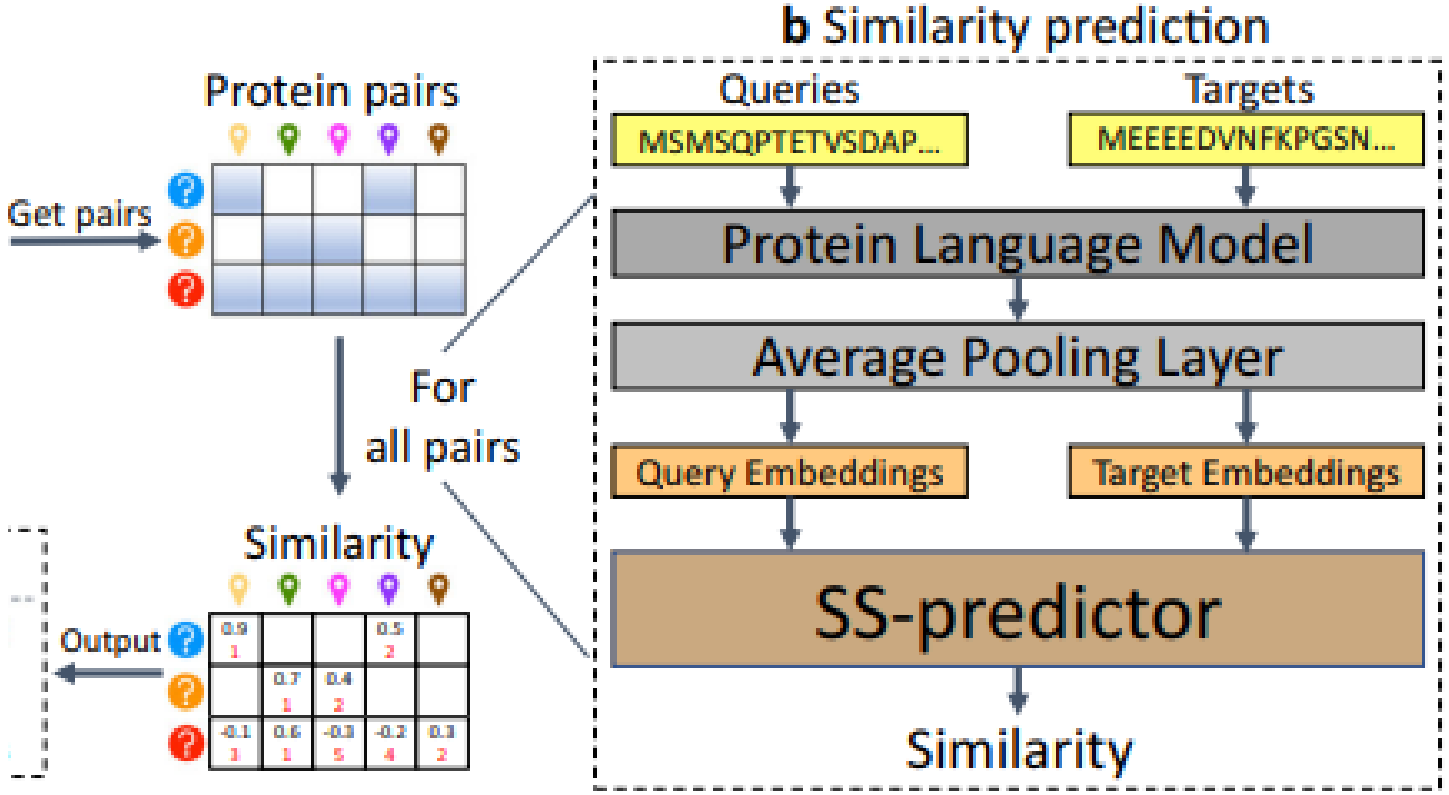

a PfamClan
Query 1
Query 2
Query 3
PfamScan
Queries
Targets
Get pairs

PfamClan

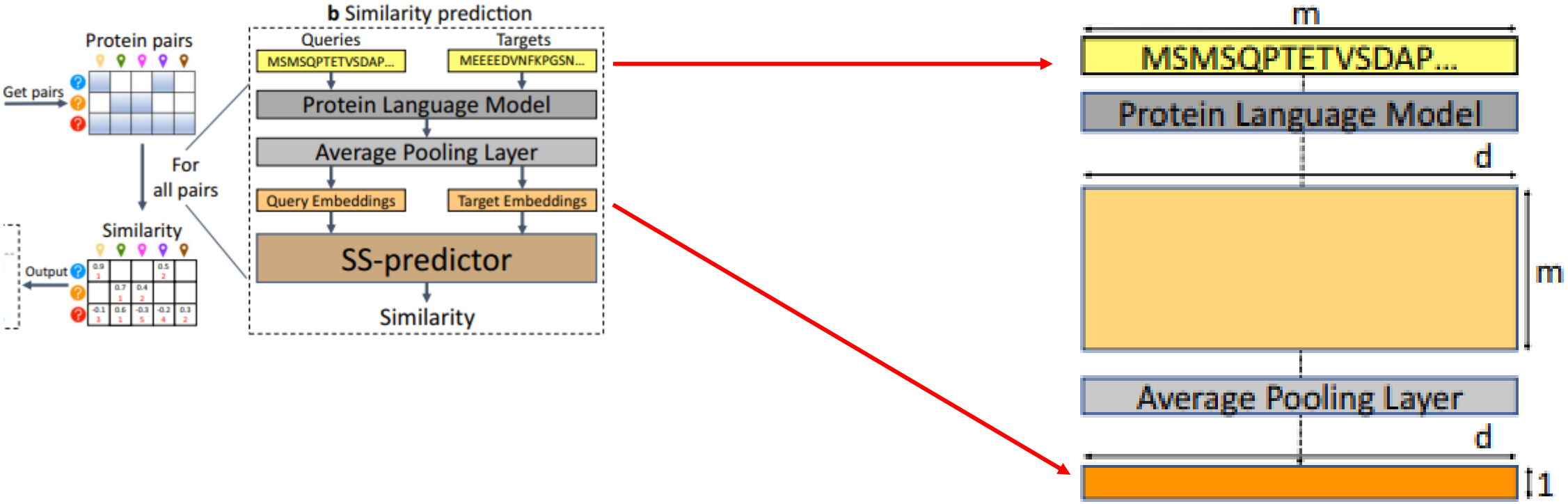

a PfamClan
Query 1
Query 2
Query 3
PfamScan
Queries
Targets
Get pairs

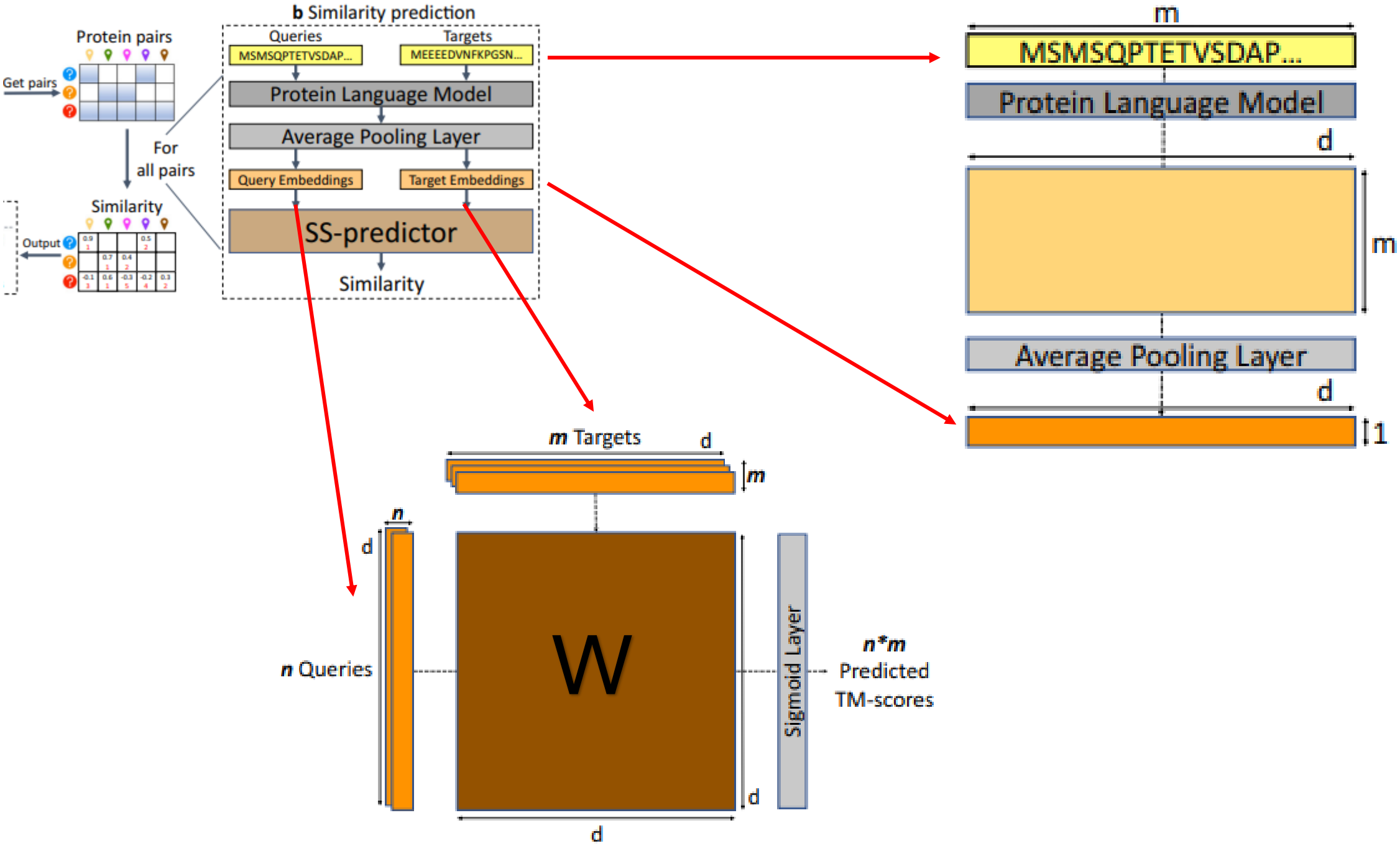
Pfam Clans
vs
Pfam Families

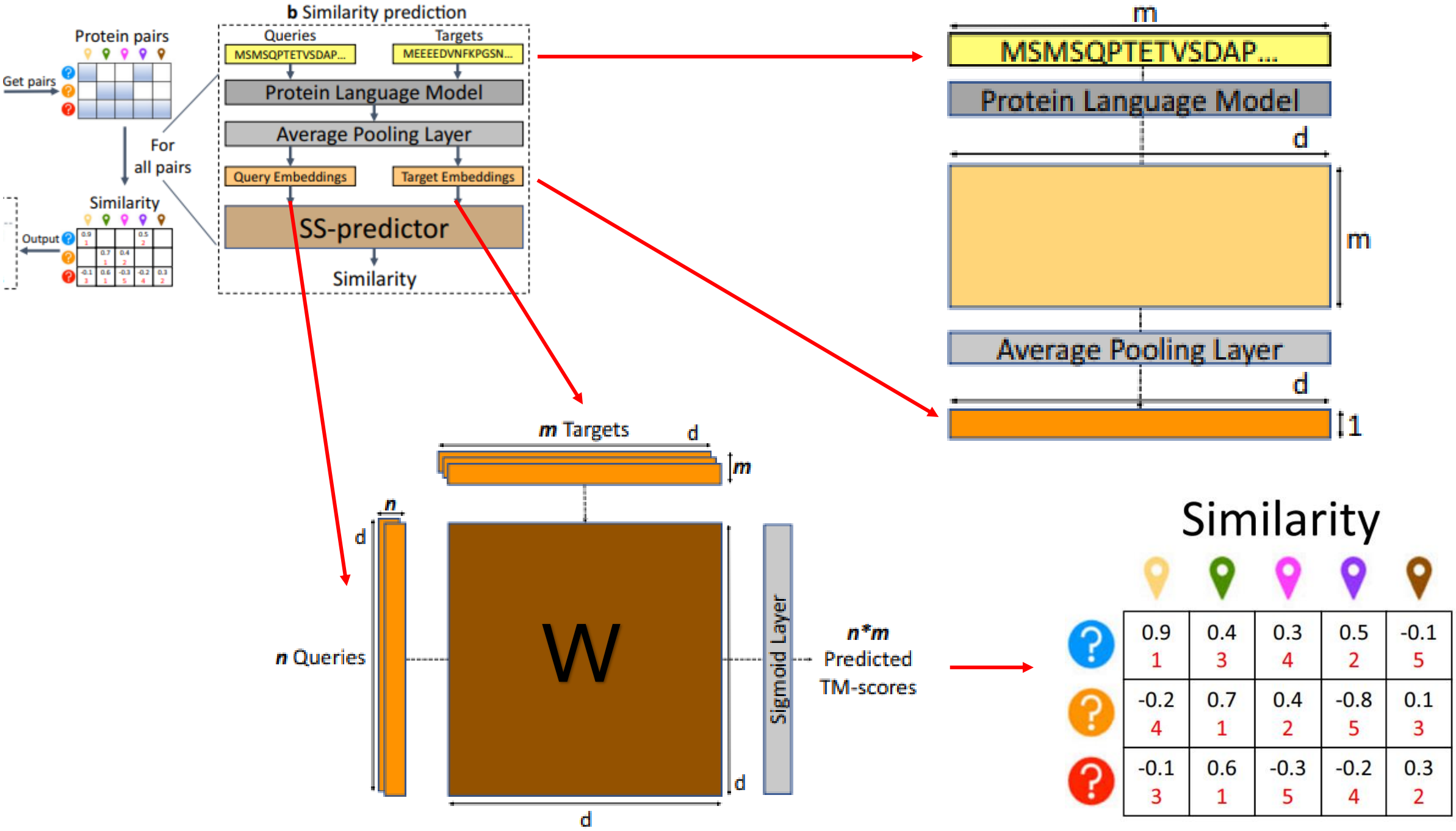




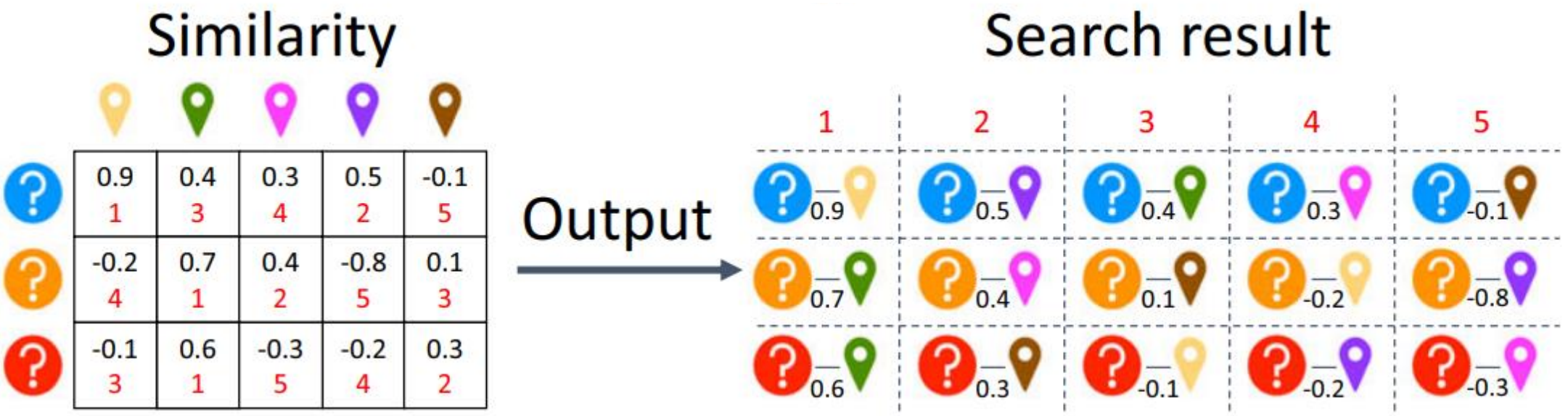
Structural Similarity Predictor (SS-predictor)


b Similarity prediction
Protein pairs
Get pairs
For
all pairs
Similarity
Output
Queries
Targets
MSMSQPTETVSDAP...
MEEEEDVNFKPGSN...
Protein Language Model
Average Pooling Layer
Query Embeddings
Target Embeddings
SS-predictor
Similarity
m
MSMSQPTETVSDAP...
Protein Language Model
d
m
Average Pooling Layer
d
1


b Similarity prediction
Protein pairs
Get pairs
For all pairs
Similarity
Output
Queries
Targets
MSMSQPTETVSDAP...
MEEEEDVNFKPGSN...
Protein Language Model
Average Pooling Layer
Query Embeddings
Target Embeddings
SS-predictor
Similarity
m
MSMSQPTETVSDAP...
Protein Language Model
d
m
Average Pooling Layer
d
1
m Targets
d
m
n
d
n Queries
W
d
d
Sigmoid Layer
n*m
Predicted
TM-scores

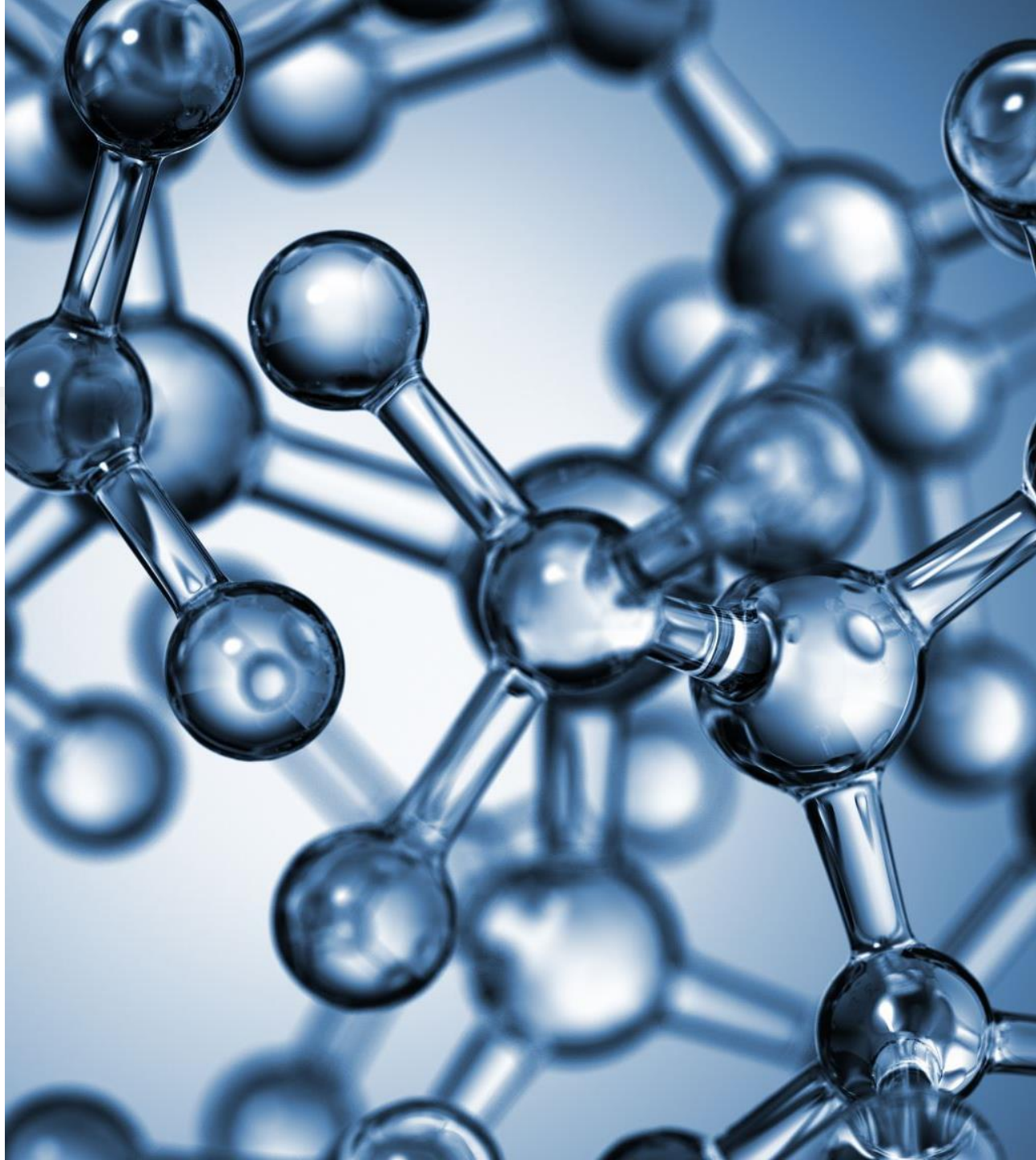

b Similarity prediction
Protein pairs
Get pairs
For all pairs
Similarity
Output
Queries
Targets
MSMSQPTETVSDAP...
MEEEEDVNFKPGSN...
Protein Language Model
Average Pooling Layer
Query Embeddings
Target Embeddings
SS-predictor
Similarity
m
MSMSQPTETVSDAP...
Protein Language Model
d
m
Average Pooling Layer
d
1
m Targets
d
m
n
d
n Queries
W
d
d
Sigmoid Layer
n*m
Predicted
TM-scores
Similarity
0.9 1
0.4 3
0.3 4
0.5 2
-0.1 5
-0.2 4
0.7 1
0.4 2
-0.8 5
0.1 3
-0.1 3
0.6 1
-0.3 5
-0.2 4
0.3 2

SS-predictor: Final Output

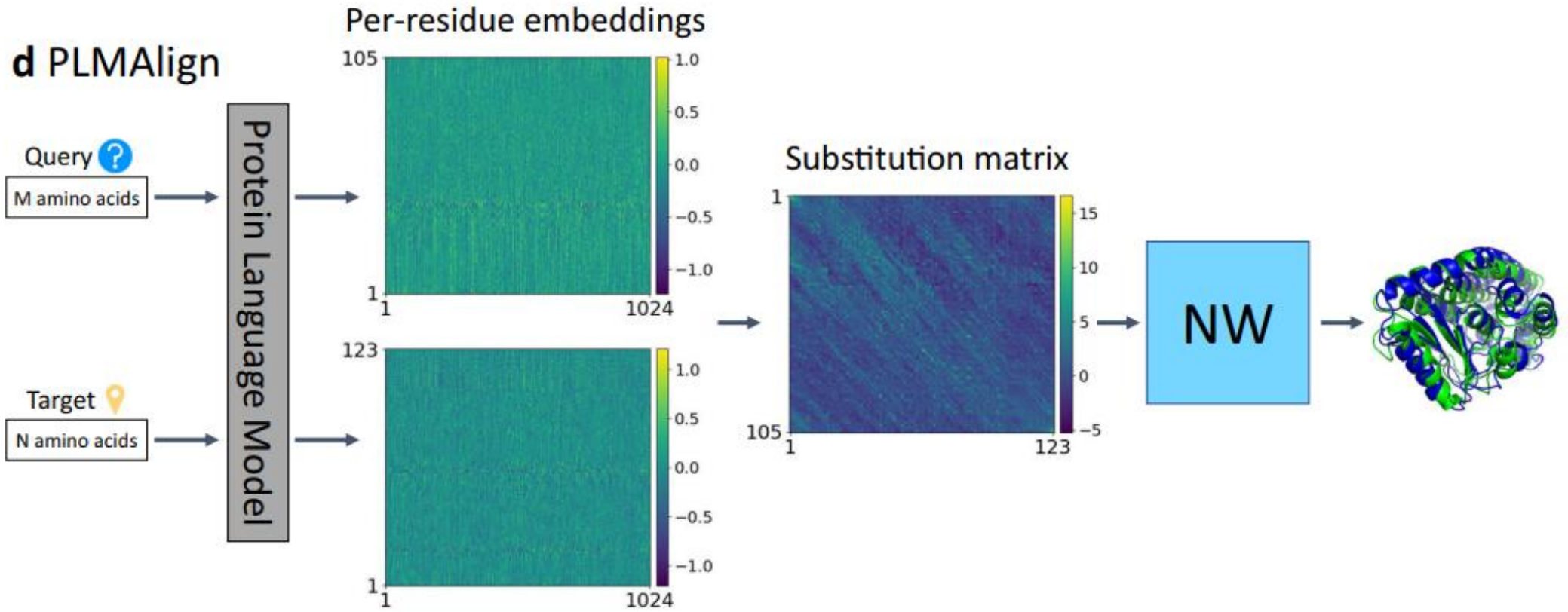


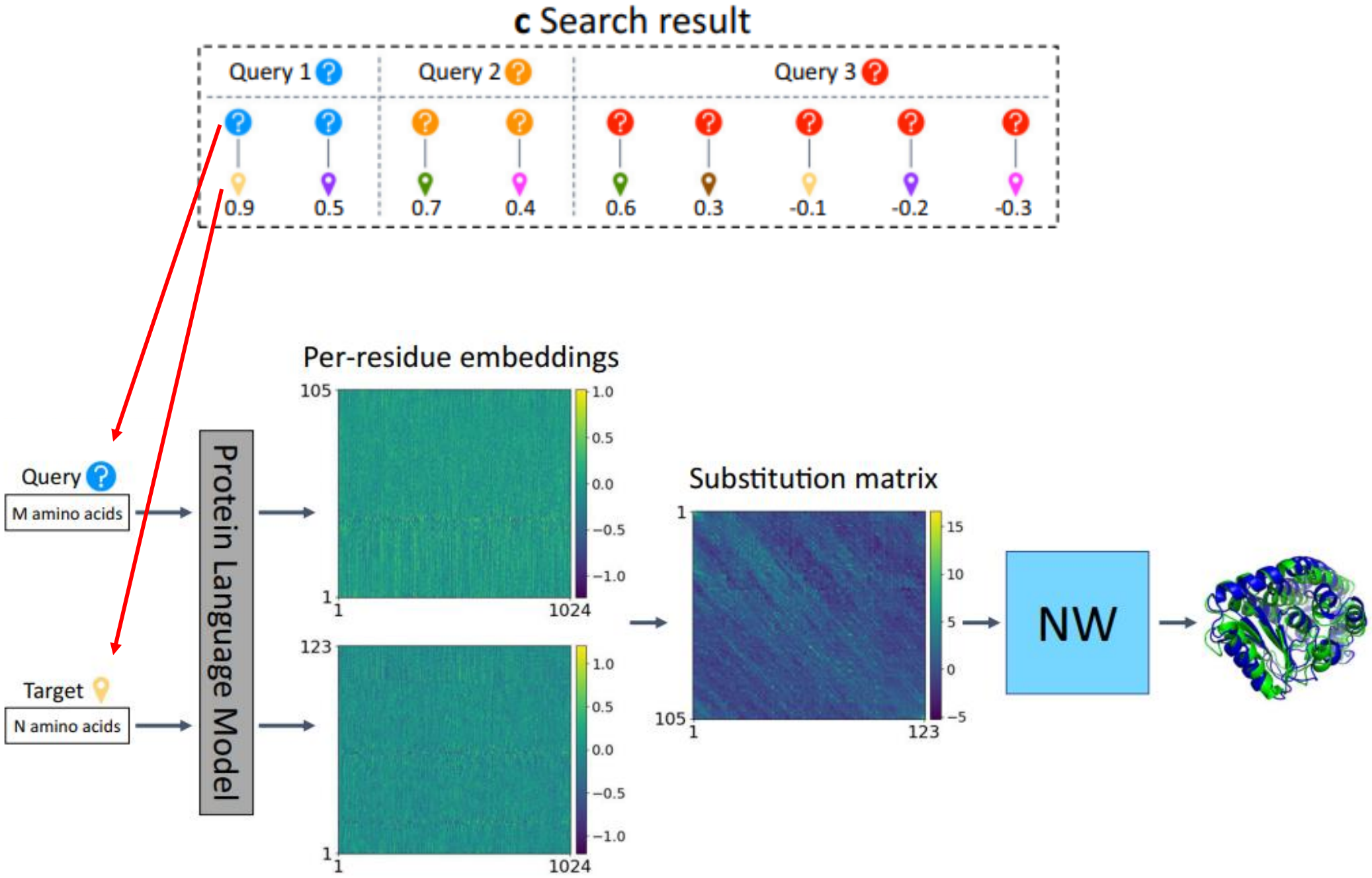
SS-predictor: Key Advantages

1) **Efficiency:** Avoids the computational cost of structure-based alignment for large datasets
2) **Scalability:** Handles the increasing size of protein clusters effectively through sequence-based predictions
3) **Precision:** Combines TM-score and COS similarity to handle both low and high sequence identity cases
4) **Global Similarity Focus:** Optimized for detecting **global structural similarity**, making it suitable for applications requiring a complete fold detection



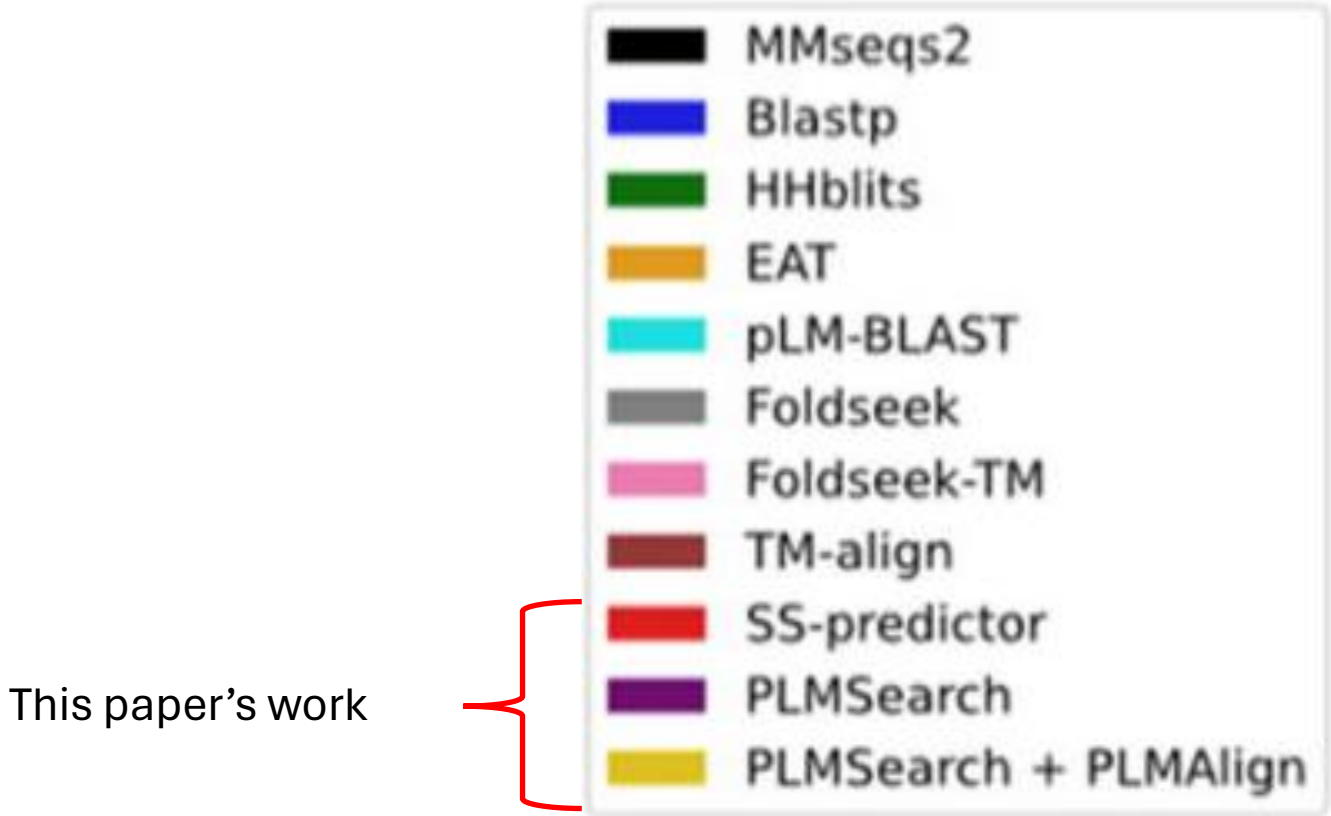
PLMAlign


d PLMAlign
Query
M amino acids
Target
N amino acids
Protein Language Model
Per-residue embeddings
105
1
1
1024
1.0
0.5
0.0
−0.5
−1.0
123
1
1
1024
1.0
0.5
0.0
−0.5
−1.0
Substitution matrix
1
105
1
123
15
10
5
0
−5
NW

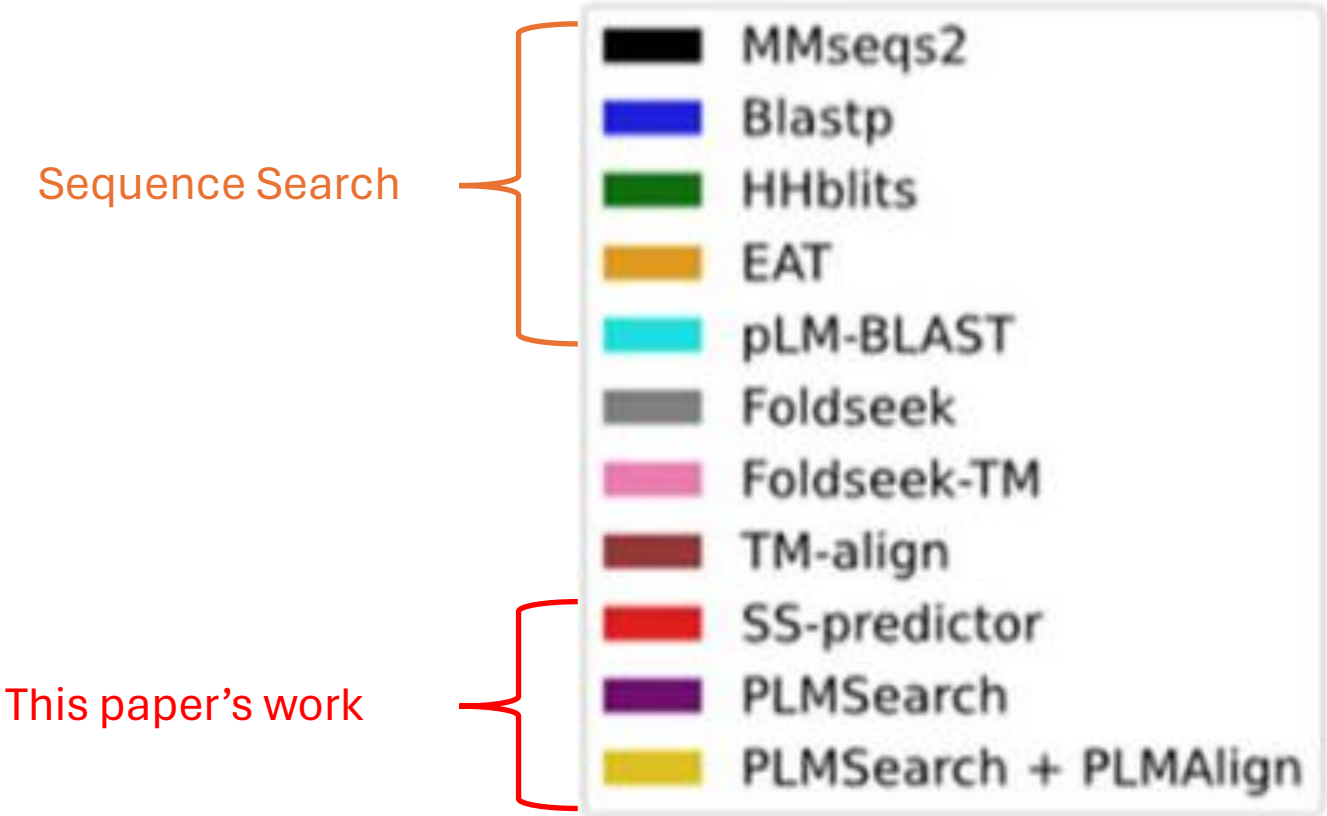

c Search result
Query 1
Query 2
Query 3
0.9
0.5
0.7
0.4
0.6
0.3
-0.1
-0.2
-0.3
Query
M amino acids
Target
N amino acids
Protein Language Model
Per-residue embeddings
105
1
1
1024
123
1
1
1024
1.0
0.5
0.0
−0.5
−1.0
Substitution matrix
1
105
1
123
15
10
5
0
−5
NW

Evaluation/Results

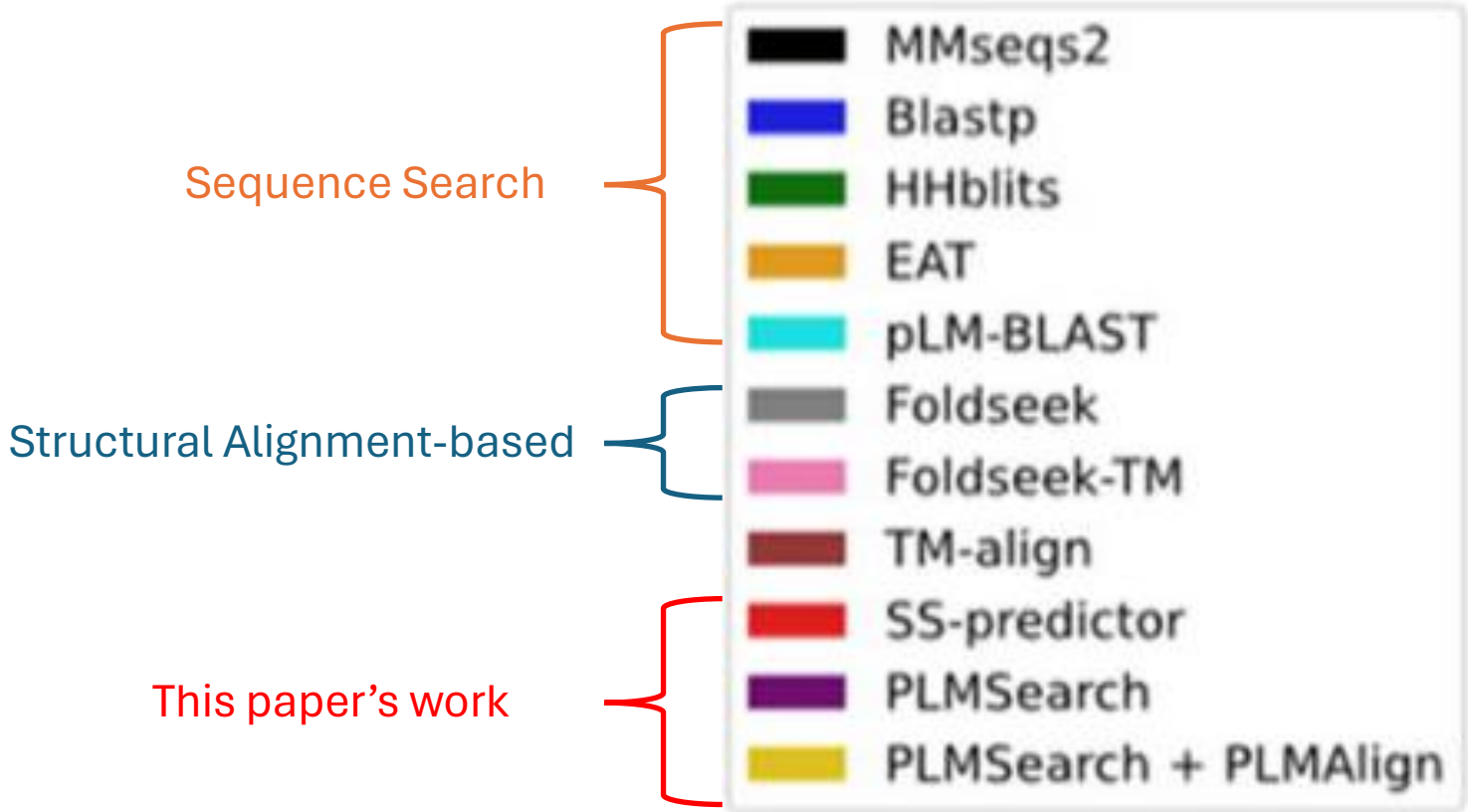
Results: Sensitivity Benchmarks



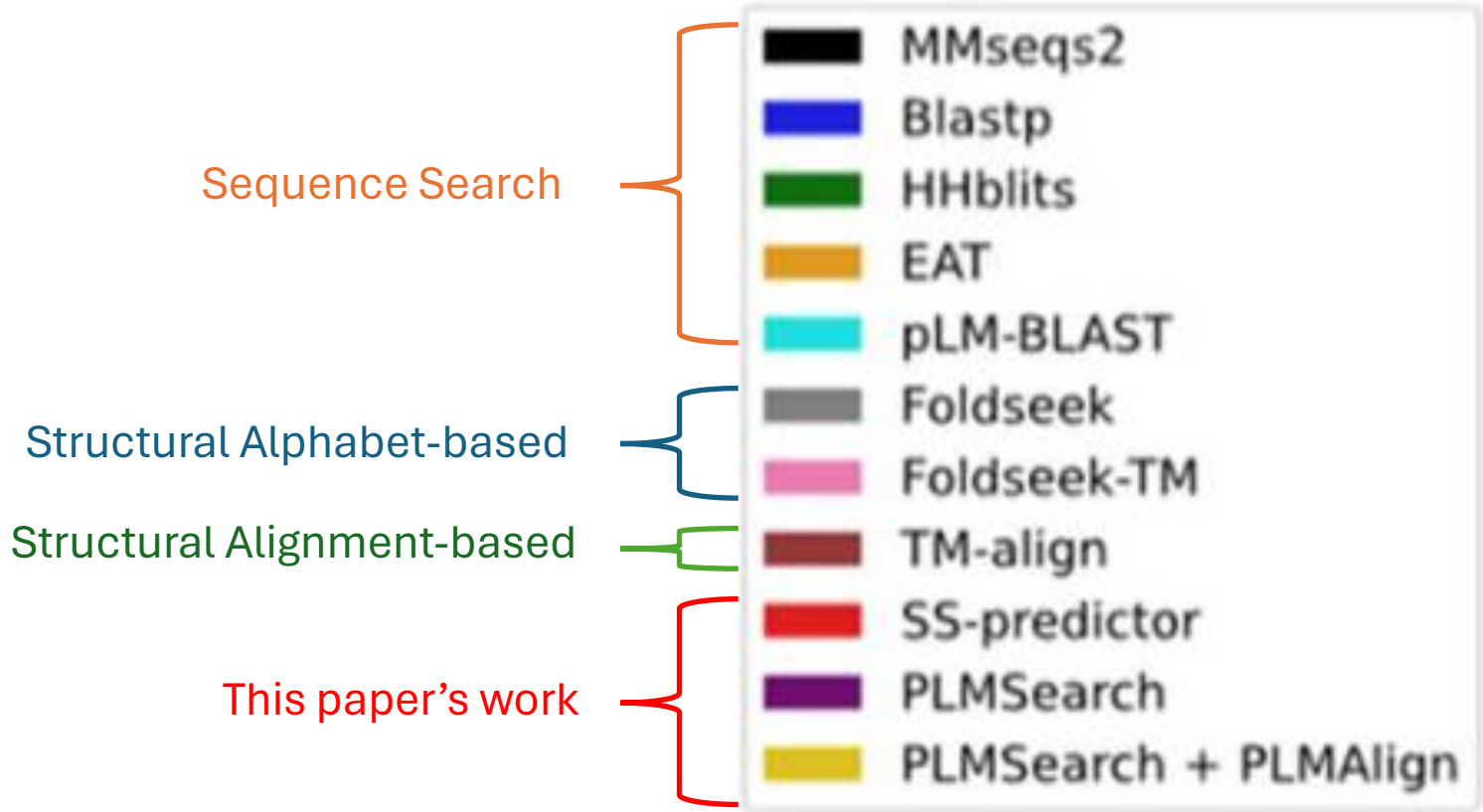
Results: Sensitivity Benchmarks



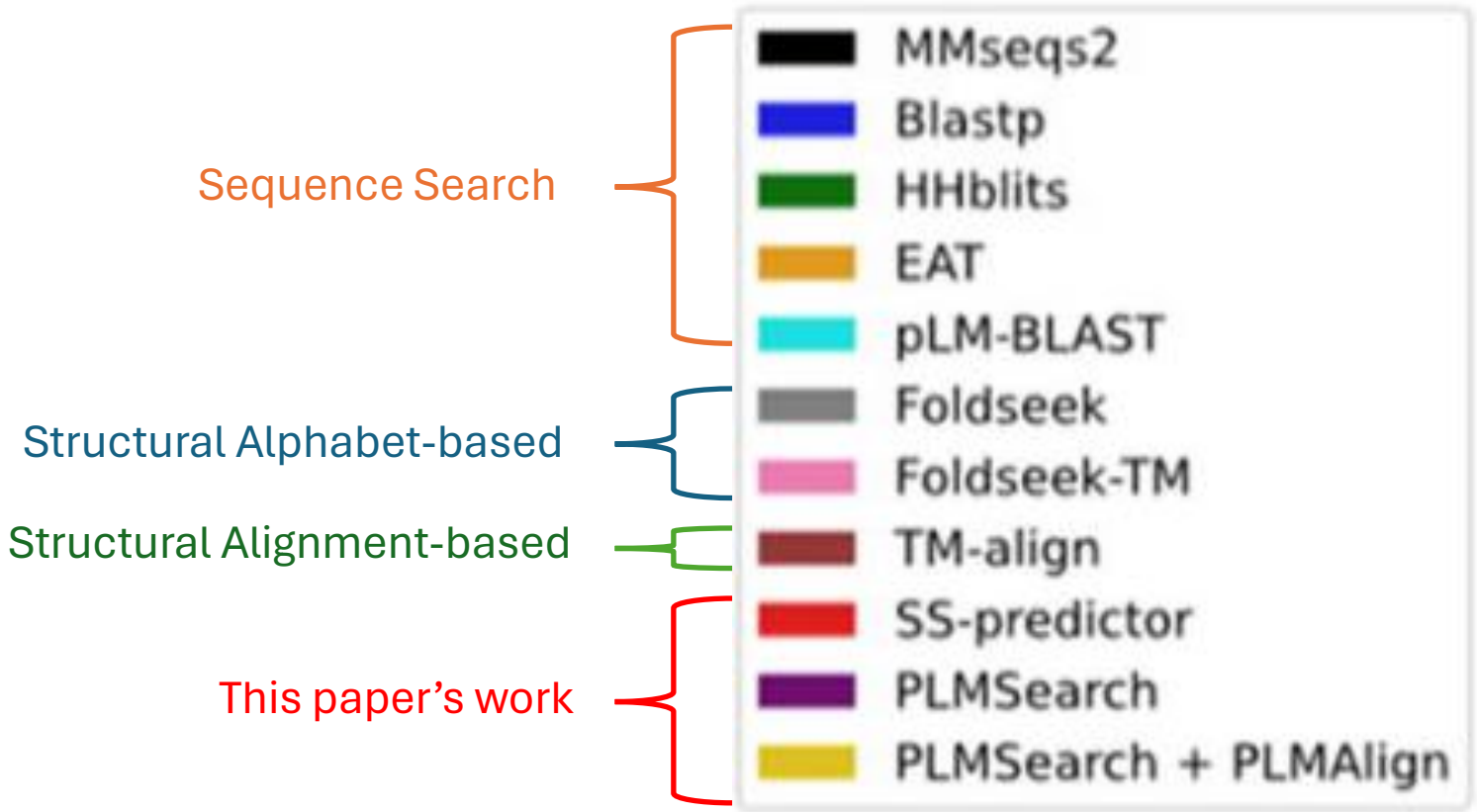
Results: Sensitivity Benchmarks



Results: Sensitivity Benchmarks

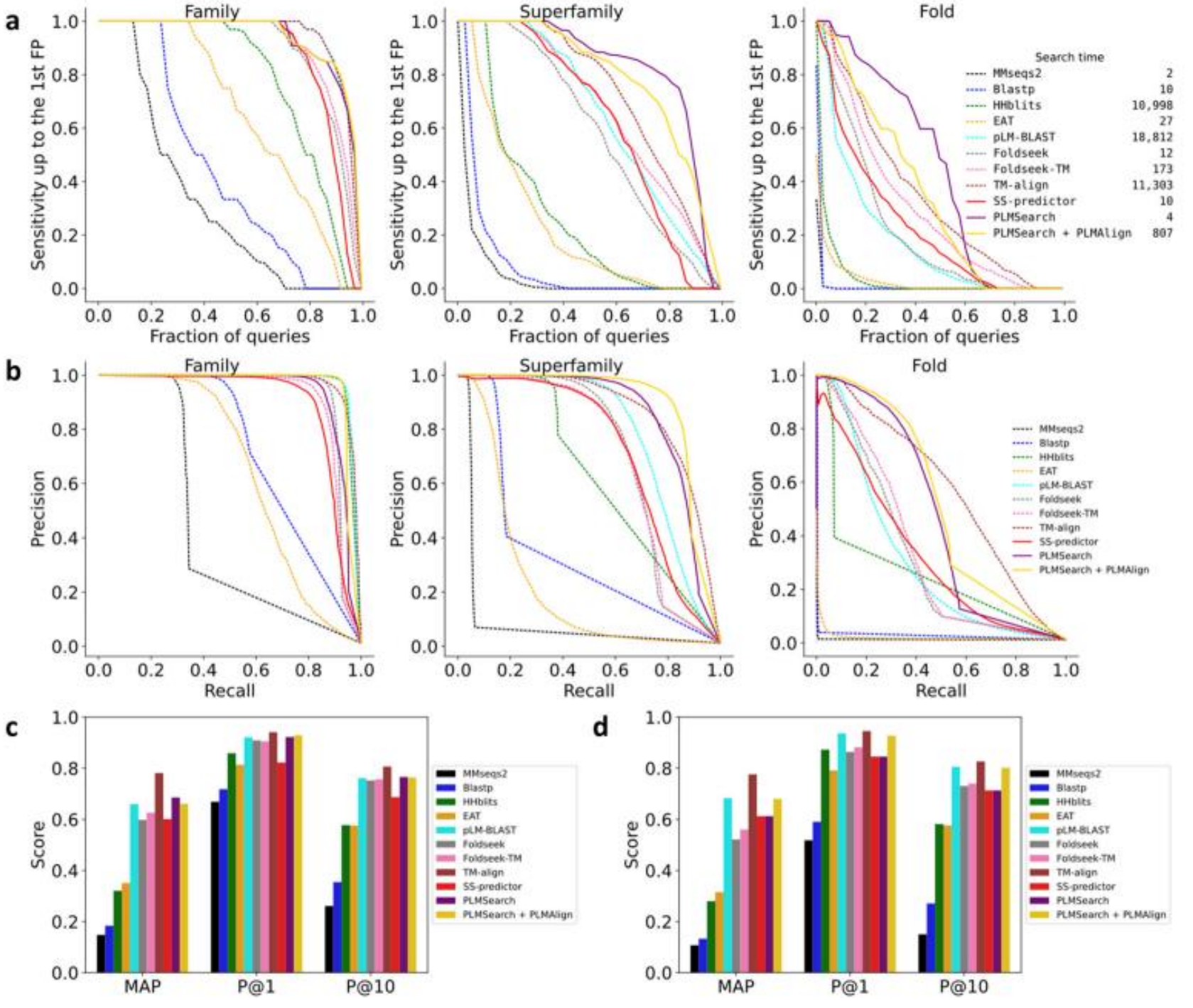


Results: Sensitivity Benchmarks

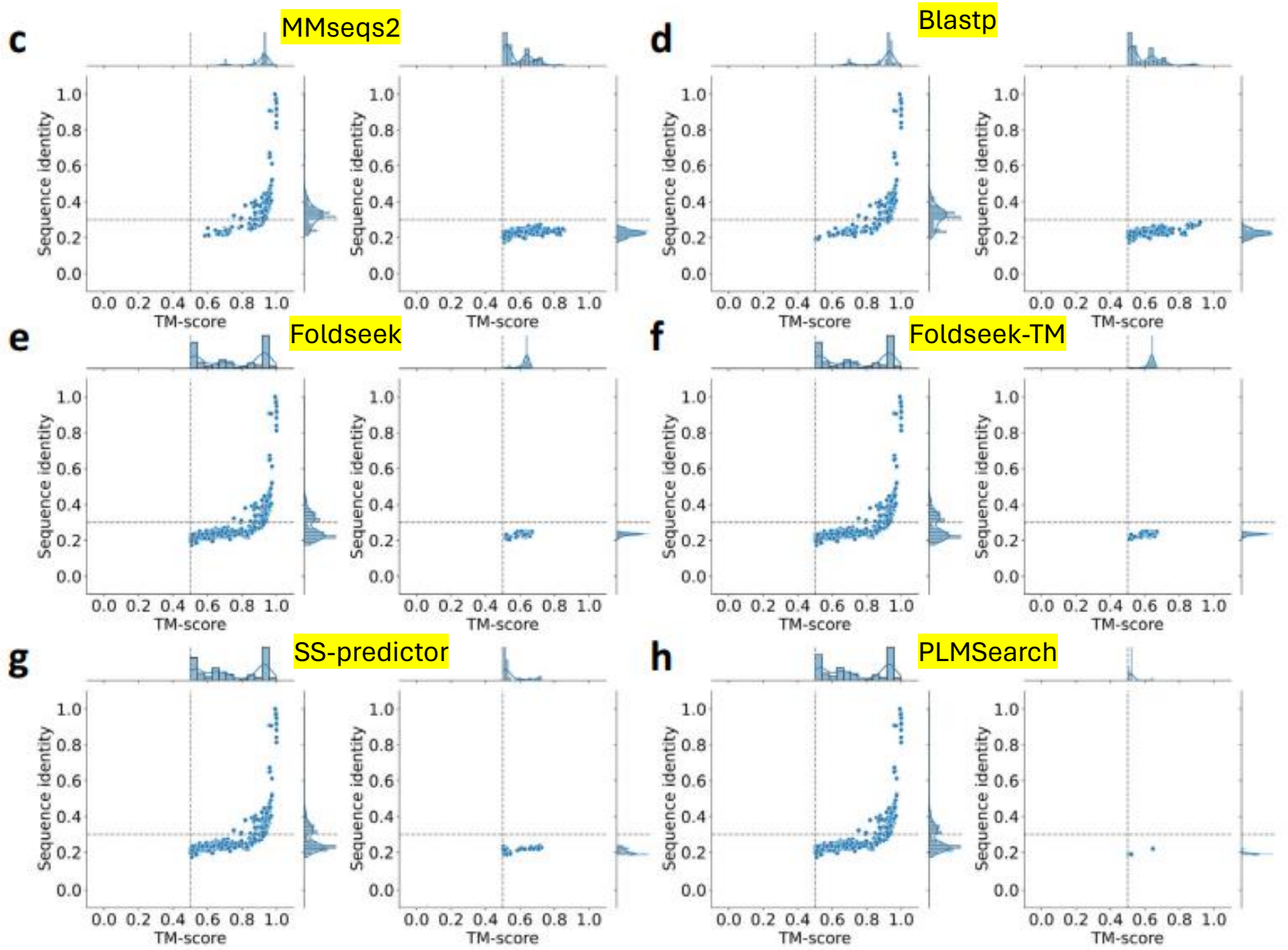


- SCOPe40-test dataset (2207 proteins)
- All-versus-all search test is performed
- 4,870,849 total query-target pairs

All-vs-All Search Test on SCOPe40-test



All-vs-All Search Test on SCOPe40-test



Results Summary

- PLMSearch is comparable to S.O.T.A. structural search methods
- PLMSearch overcame low sensitivity cap on sequence methods
- Improvements are concentrated in remote homology pairs
- PLMSearch is one of the fastest search methods
- Residue embedding-based alignment methods are limited by size of target dataset

Future Works

Efficient Embedding Storage

Thank You!


Questions?