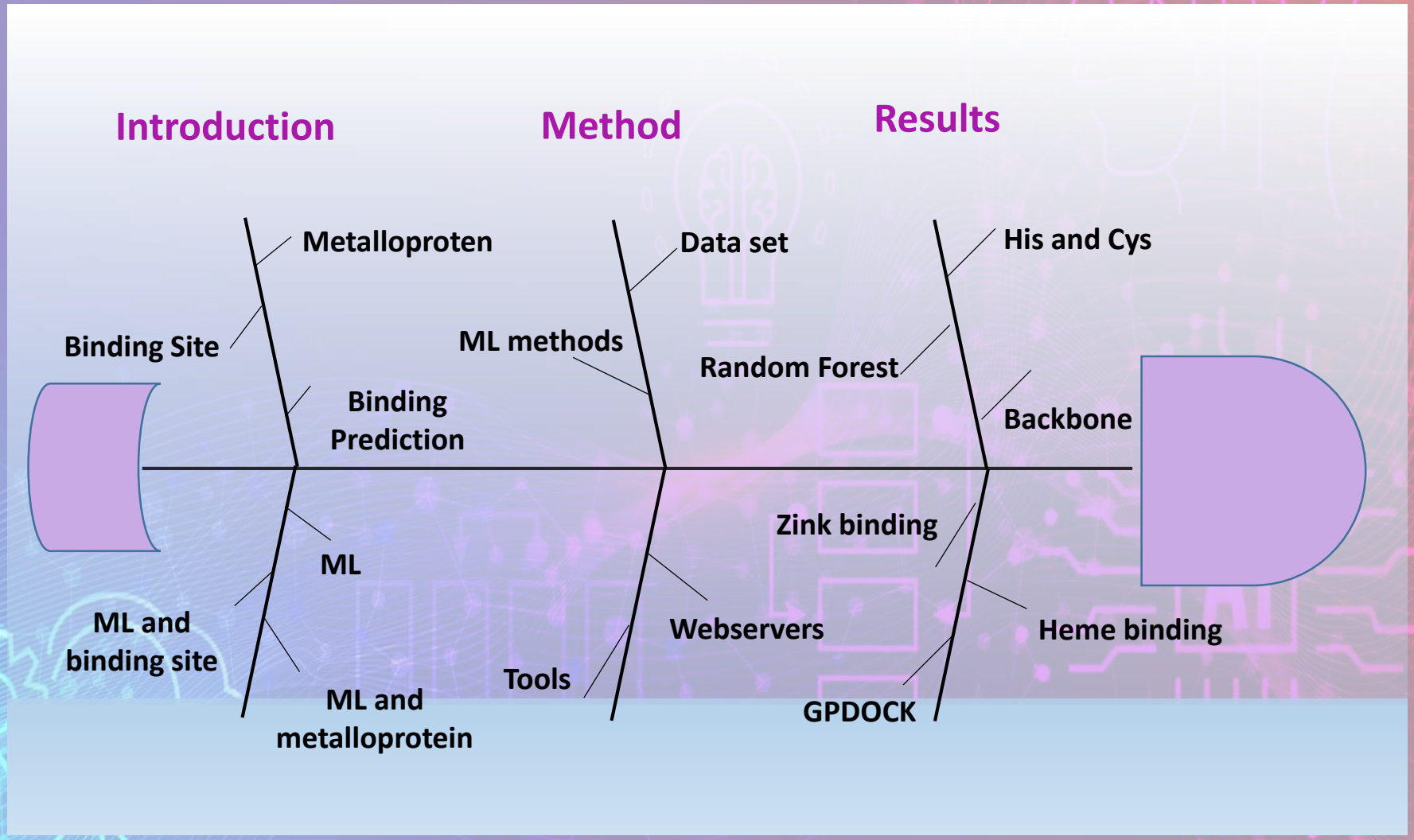
Predicting Metalloprotein Binding Sites: A Machine Learning Approach

Supervisor: Dr. Kaveh Kavousi

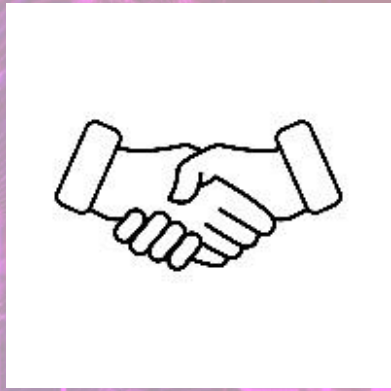
Presenter : Fereshteh Noroozi

2024

Table of content

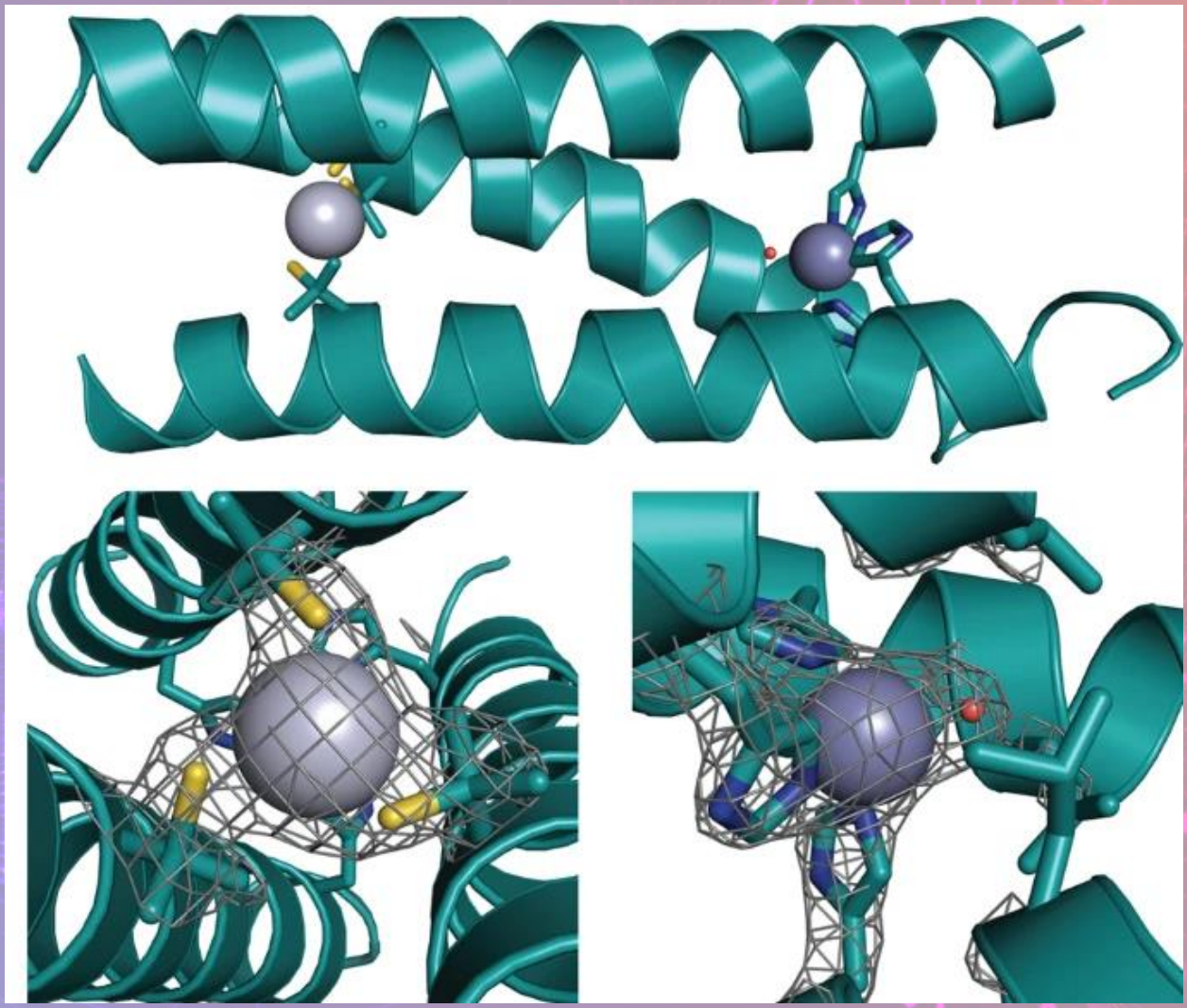


Introduction



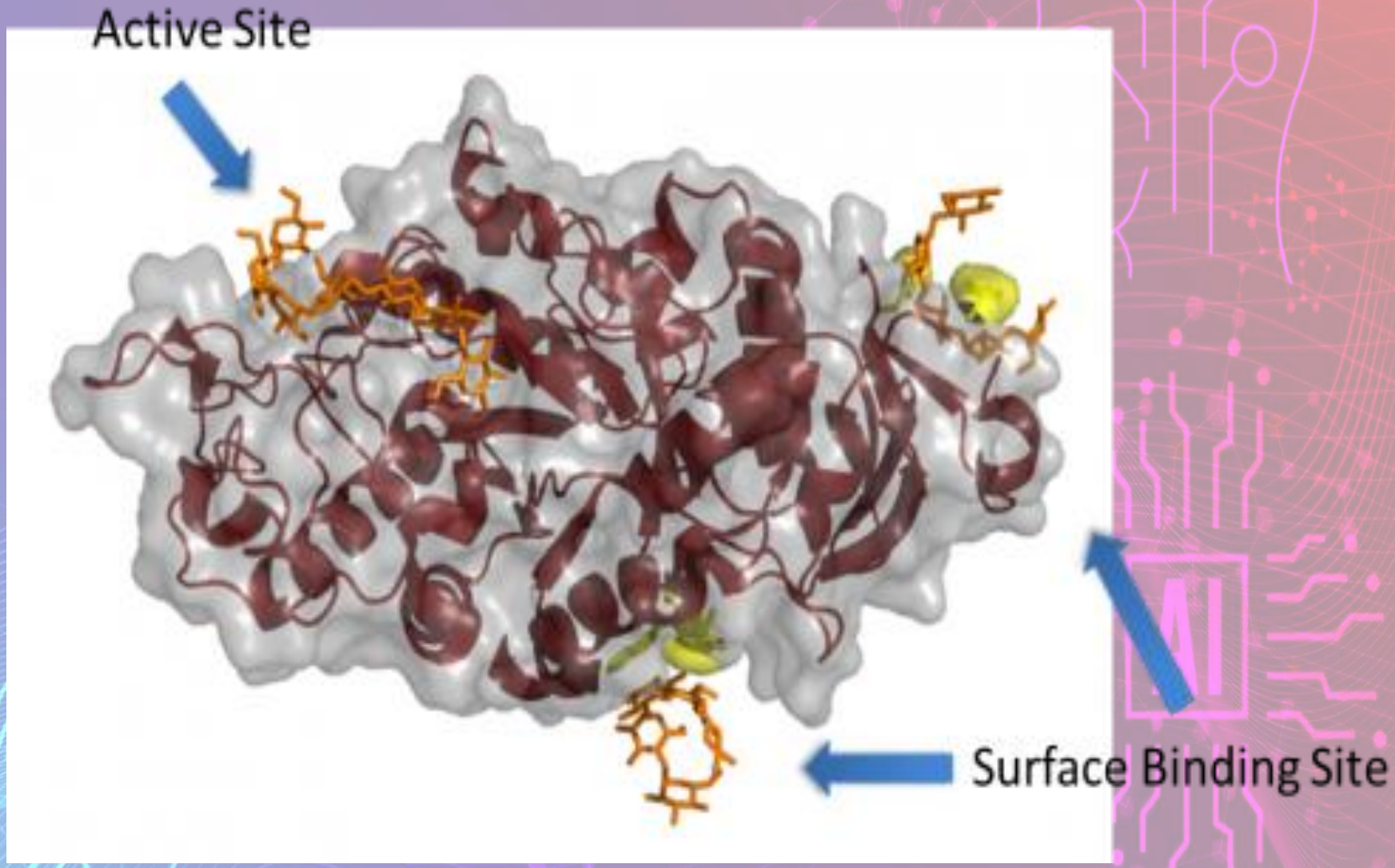
Metalloproteins

- Metalloproteins are a class of proteins that contain metal ions as essential components for their structure and function



Hydrolytic catalysis and structural stabilization in a designed metalloprotein, Nature Chemistry, 2012

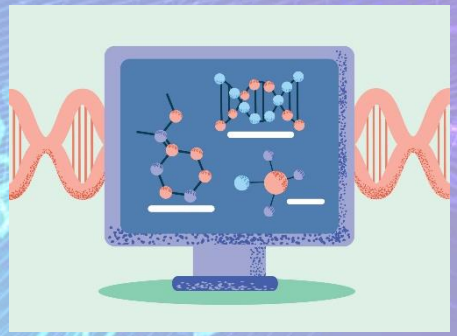
Crucial Binding Site Significance

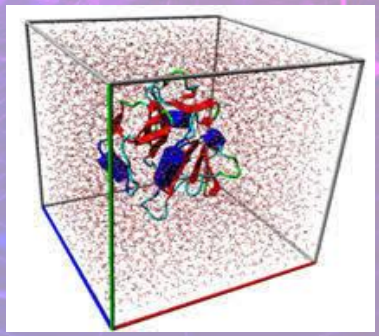


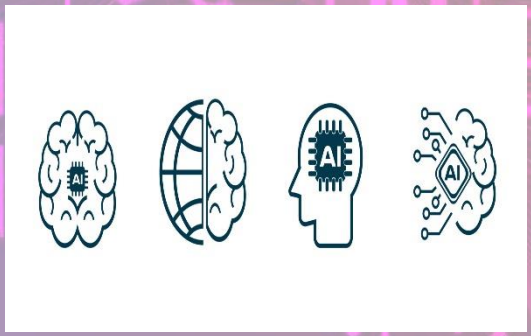
PDB ID 1rp8-alpha-amylase isozyme 1

Diverse Methods for Binding Sites Prediction

1. • **Sequence-Based Methods**
2. • **Structure-Based Methods**
3. • **AI Approaches**

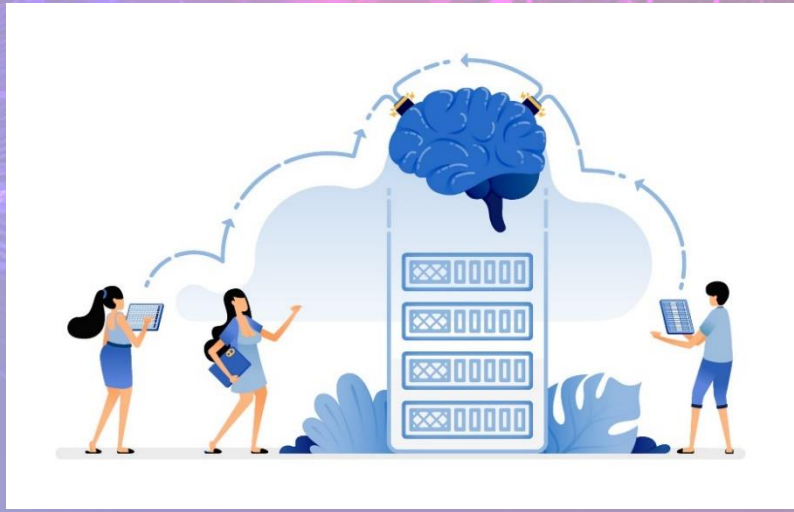






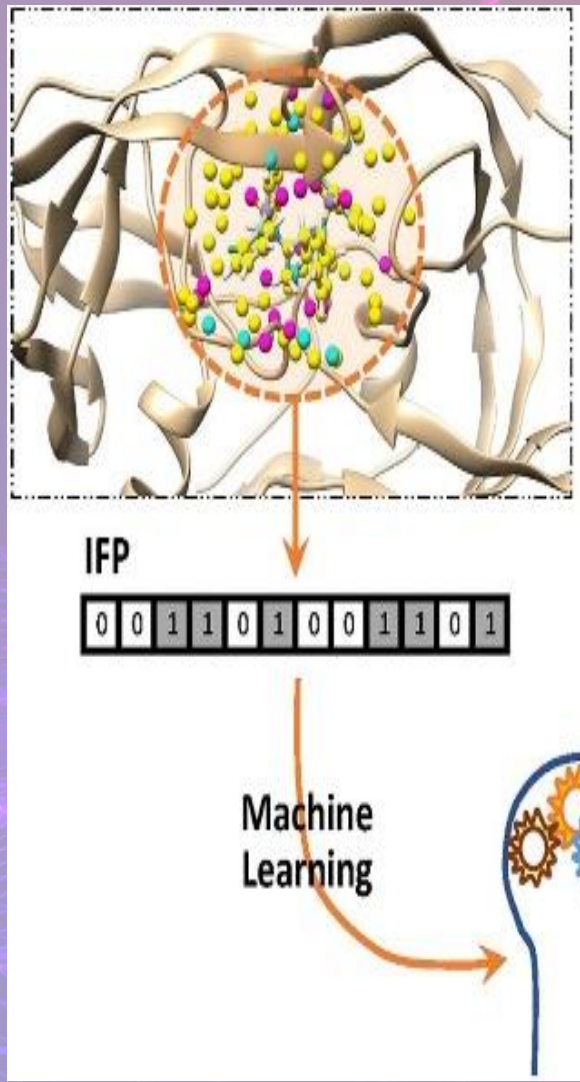
Discovering with Machine Learning

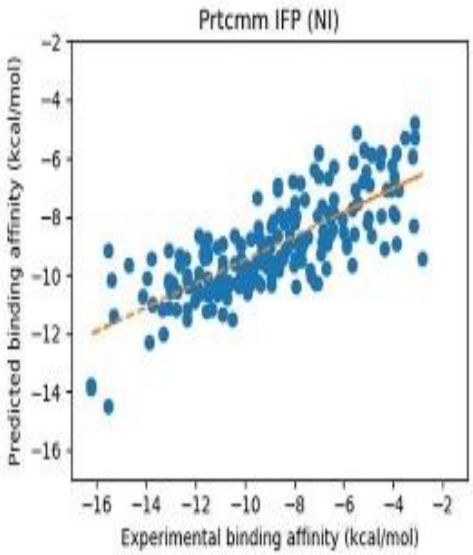
1. • Autonomous extraction of influential patterns.
2. • Significance in Data Science for precise estimations.
3. • Empowering data scientists through data-driven predictions.
4. • Emerging Technologies



Utilizing Machine Learning for Prediction of Metalloprotein Binding Sites

- Naive Bayesian
- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine
- Metalloprotein Prediction





Structure-based protein–ligand interaction fingerprints for binding affinity predictionComputational and Structural Biotechnology Journal,2021

Purpose

Research Interest

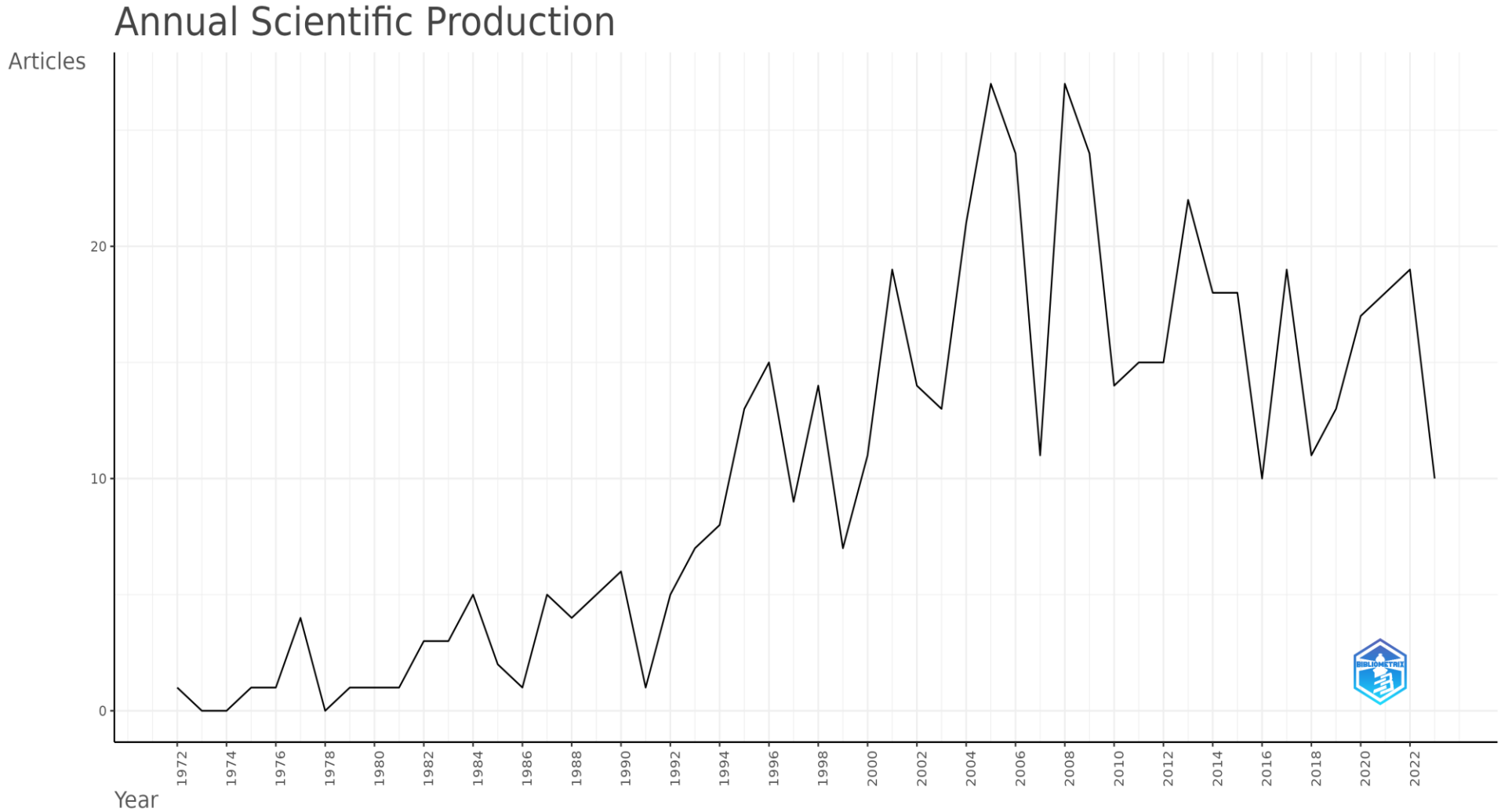
Explore ML for Metalloprotein Binding

Investigate Efficacy

Address Ion Challenges in Binding Pockets

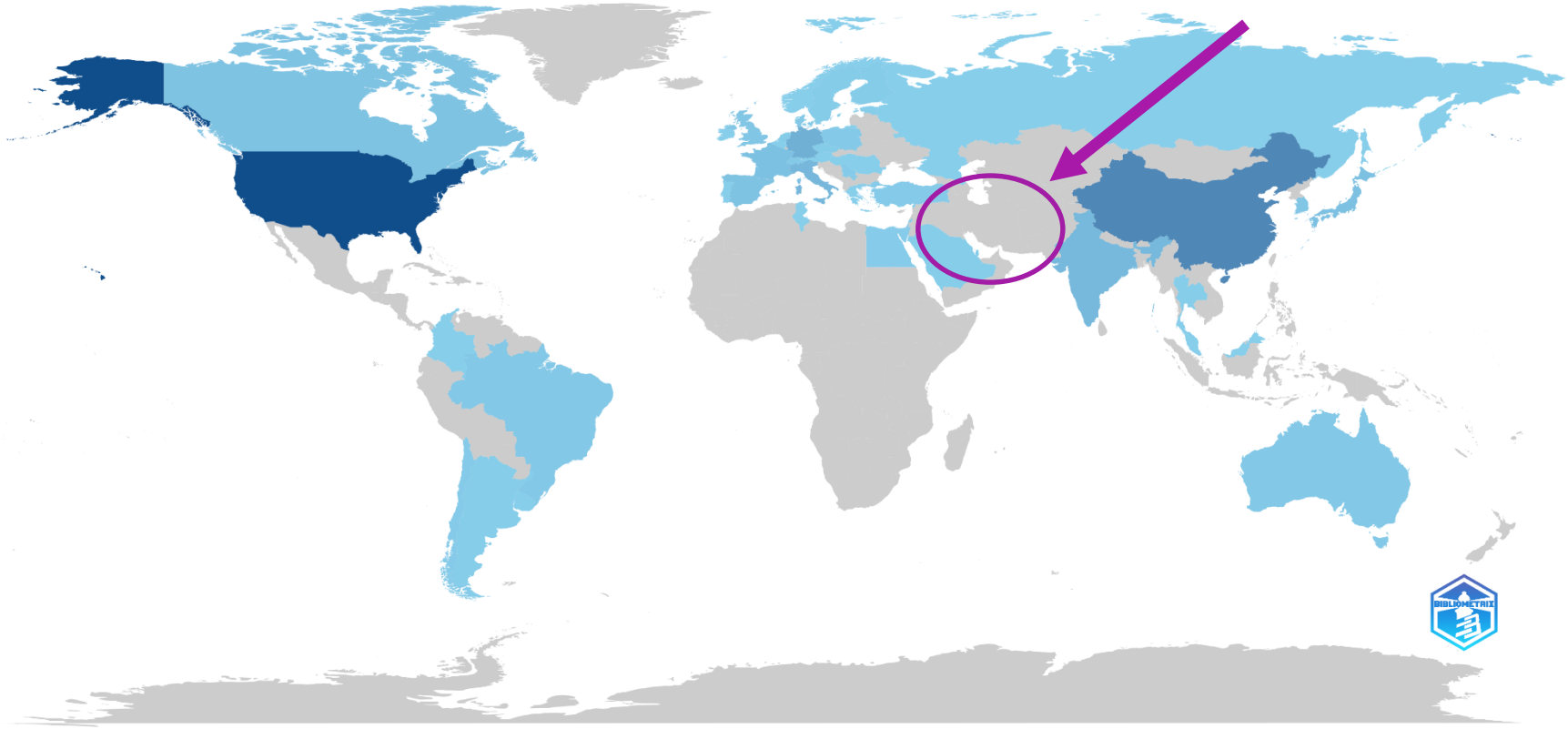
Amalgamate Research & Guide Responses

Topic Statistics



Topic Statistics

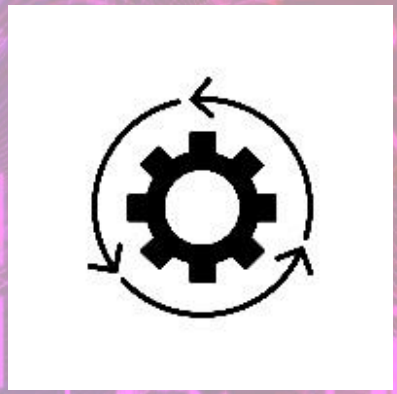
Country Scientific Production



Research background

Paper Title	Year	Model
Identifying Cysteines and Histidines in Transition-Metal-Binding Sites Using Support Vector Machines and Neural Networks	2006	SVM
Predicting zinc binding at the proteome level	2007	SVM
SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides	2014	SVM
Prediction of Metal Ion Binding Sites in Proteins from Amino Acid Sequences by Using Simplified Amino Acid Alphabets and Random Forest Model	2017	Random Forest Model
Identifying metal binding amino acids based on backbone geometries as a tool for metalloprotein engineering	2021	Random Forest Model
GPDOCK: highly accurate docking strategy for metalloproteins based on geometric probability	2023	logistical regression model

Methods



Data Collection

Initial Data Collection	Data Filtering	Final Data
PDB Uniprot MetalPDB PROSITE	2.5 Å Res HSSP value cd-hit UniRef 50 S 90%	Clustering Annotation PyMOL MBS Balancing the Dataset Selection of Ligands

Feature Selection

- Position-specific evolutionary profiles
- Global descriptor
- Conservation Features
- Conformational Similarity
- BLOSUM 50 Substitution Matrix
- Hydrophobicity
- Amino Acid Composition (AAC)
- Dipeptide Composition (DPC)

$$\text{Sequence Length Relative to Average} = \frac{L-\bar{L}}{\bar{L}}$$

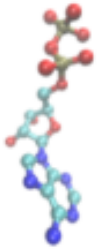
Where:

- L represents the sequence length of the protein chain being considered.
- $\bar{L}$ represents the average sequence length of all protein chains in the training set.

PSI-BLAST	Conservation of CYS	Conservation of HIS
4	00100	00010
3	01000	10000
0	00010	00001
2	10000	00100
1	00001	01000

Feature Selection

(a). Atomic sets

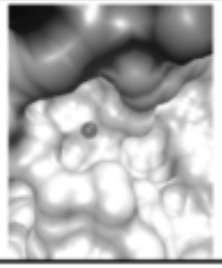


α: Heavy atoms of Ligand

γ: O/N/S in Ligand

ε: Not O/N/S in Ligand

η: Metal ions in Ligand



β: Heavy atoms of Protein

δ: O/N/S in Protein

ζ: Not O/N/S in Protein

θ: Metal ion

(b). 36 features

$\sum_i^{\alpha}\sum_j^{\beta}(d_{ij} < 3.5)$	$\sum_i^{\alpha}\sum_j^{\beta}(d_{ij} < 4.5)$	$\sum_i^{\alpha}\sum_j^{\beta}(d_{ij} < 5.5)$	$\sum_i^{\gamma}\sum_j^{\delta}(d_{ij} < 3.5)$	$\sum_i^{\gamma}\sum_j^{\delta}(d_{ij} < 4.0)$	$\sum_i^{\gamma}\sum_j^{\delta}(d_{ij} < 4.5)$
$\sum_i^{\gamma}\sum_j^{\delta}(d_{ij} < 5.0)$	$\sum_i^{\gamma}\sum_j^{\delta}(d_{ij} < 5.5)$	$\sum_i^{\gamma}\sum_j^{\zeta}(d_{ij} < 3.5)$	$\sum_i^{\gamma}\sum_j^{\zeta}(d_{ij} < 4.0)$	$\sum_i^{\gamma}\sum_j^{\zeta}(d_{ij} < 4.5)$	$\sum_i^{\gamma}\sum_j^{\zeta}(d_{ij} < 5.0)$
$\sum_i^{\gamma}\sum_j^{\zeta}(d_{ij} < 5.5)$	$\sum_i^{\varepsilon}\sum_j^{\delta}(d_{ij} < 3.5)$	$\sum_i^{\varepsilon}\sum_j^{\delta}(d_{ij} < 4.0)$	$\sum_i^{\varepsilon}\sum_j^{\delta}(d_{ij} < 4.5)$	$\sum_i^{\varepsilon}\sum_j^{\delta}(d_{ij} < 5.0)$	$\sum_i^{\varepsilon}\sum_j^{\delta}(d_{ij} < 5.5)$
$\sum_i^{\varepsilon}\sum_j^{\zeta}(d_{ij} < 3.5)$	$\sum_i^{\varepsilon}\sum_j^{\zeta}(d_{ij} < 4.0)$	$\sum_i^{\varepsilon}\sum_j^{\zeta}(d_{ij} < 4.5)$	$\sum_i^{\varepsilon}\sum_j^{\zeta}(d_{ij} < 5.0)$	$\sum_i^{\varepsilon}\sum_j^{\zeta}(d_{ij} < 5.5)$	$\sum_i^{\gamma}\sum_j^{\theta}(d_{ij} < 3.0)$
$\sum_i^{\gamma}\sum_j^{\theta}(d_{ij} < 3.5)$	$\sum_i^{\gamma}\sum_j^{\theta}(d_{ij} < 4.0)$	$\sum_i^{\gamma}\sum_j^{\theta}(d_{ij} < 4.5)$	$\sum_i^{\varepsilon}\sum_j^{\theta}(d_{ij} < 4.0)$	$\sum_i^{\varepsilon}\sum_j^{\theta}(d_{ij} < 4.5)$	$\sum_i^{\varepsilon}\sum_j^{\theta}(d_{ij} < 5.0)$
$\sum_i^{\varepsilon}\sum_j^{\theta}(d_{ij} < 5.5)$	$\sum_i^{\eta}\sum_j^{\delta}(d_{ij} < 5.5)$	$\sum_i^{\gamma}\sum_j^{\delta}(min_{ij})$	$\sum_i^{\gamma}\sum_j^{\zeta}(min_{ij})$	$\sum_i^{\varepsilon}\sum_j^{\delta}(min_{ij})$	$\sum_i^{\varepsilon}\sum_j^{\zeta}(min_{ij})$

GPDOCK: highly accurate docking strategy for metalloproteins based on geometric probability, briefing in bioinformatics, 2023

Model Selection-Support Vector Machine(SVM)

- **Gaussian kernels(**radial basis function (RBF) kernel**)**

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

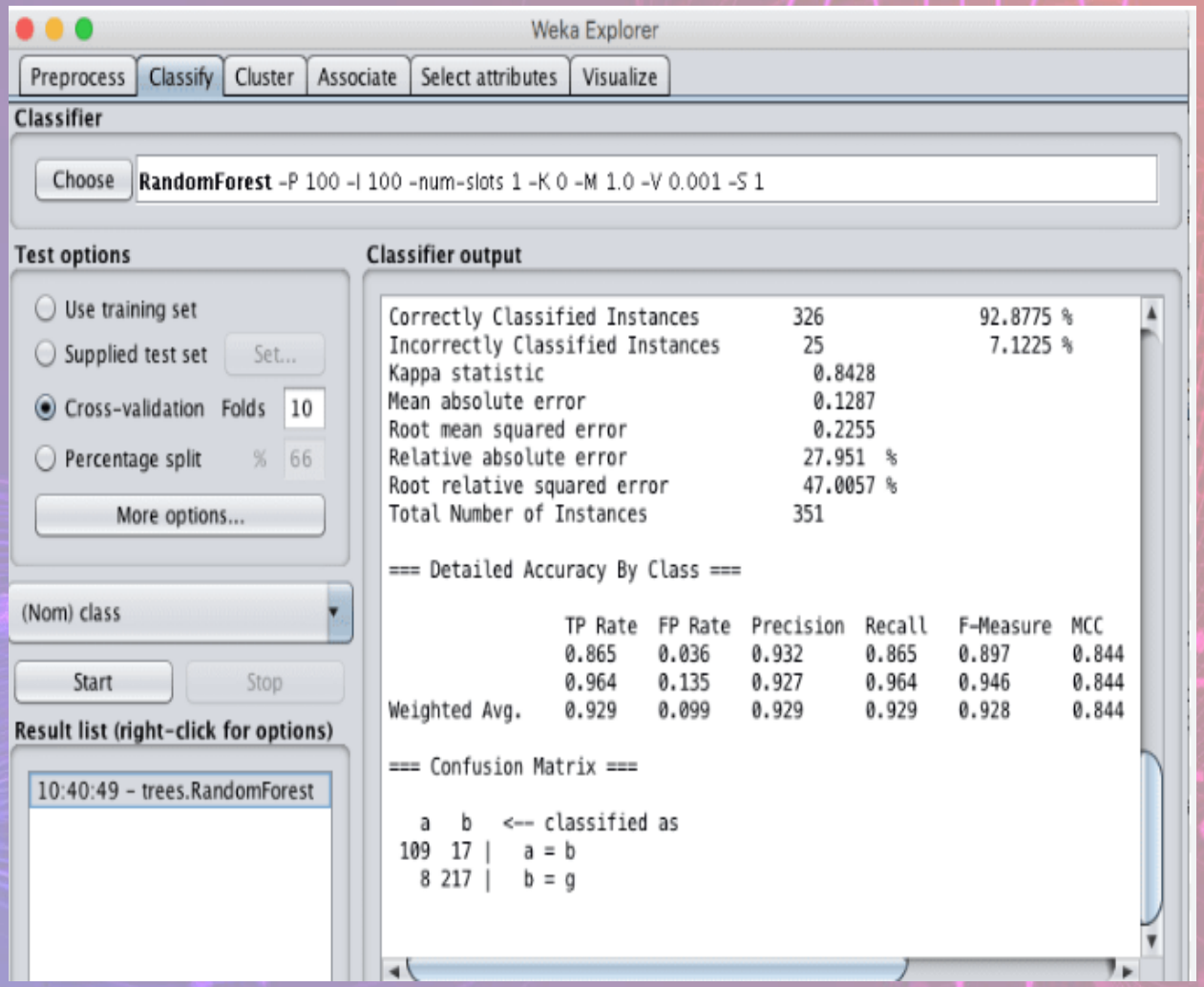
Width of the Gaussian distribution Euclidean distance

- **C regularization parameter**
- **K-fold cross validation**
- **LIBSVM package**

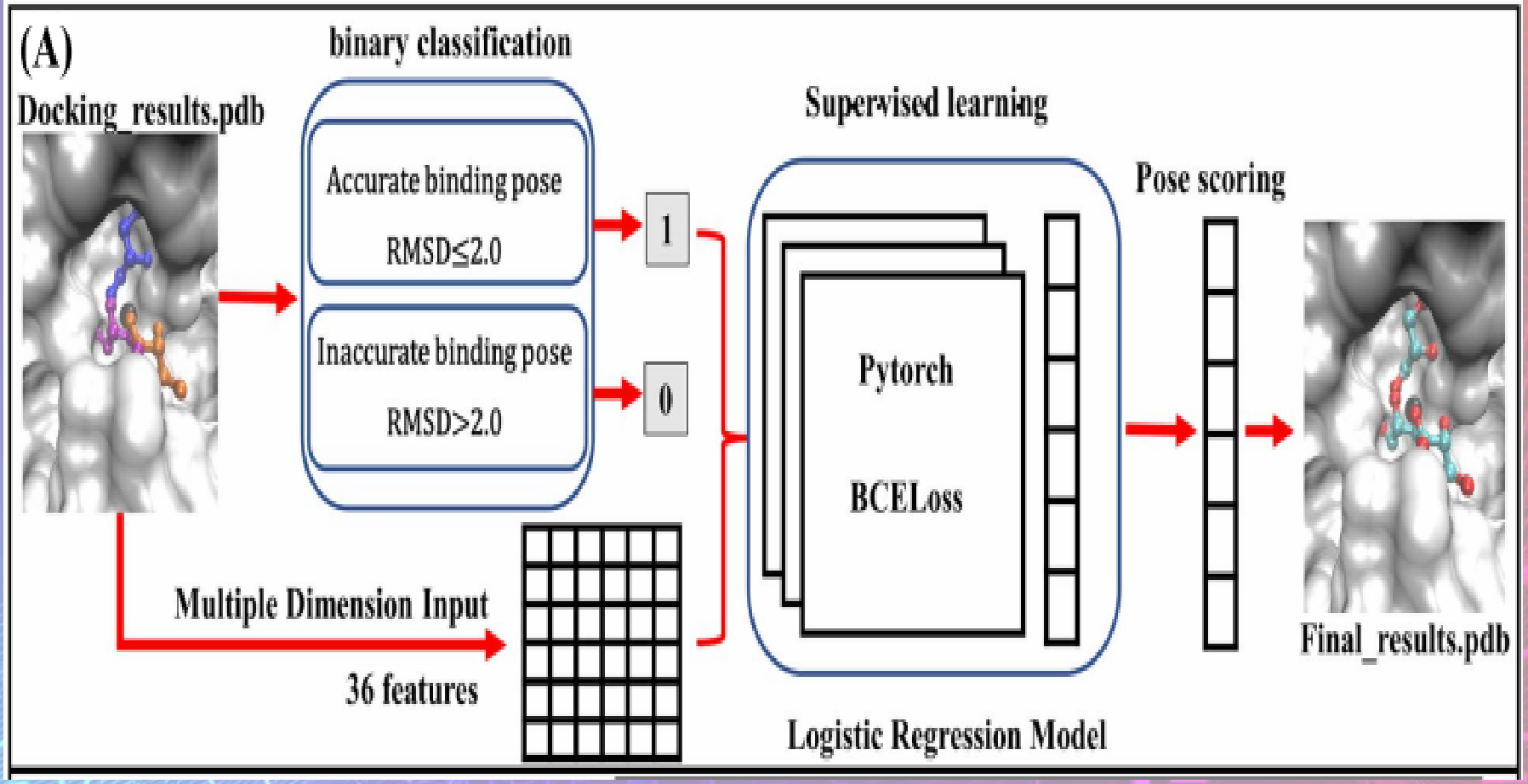
Model	g	C
Binary SVM	0.05	0.1
Multiclass SVM	0.05	5

Model Selection-Random forest consists of multiple decision trees

- **Weka package**
- **Nested cross-validation(**double cross-validation**)**
- **K-fold**
- **GridSearchCV(**scikit-learn in Python)



Model Selection-Logistic regression model



$$P(y=1|X) = \frac{1}{1+e^{-z}}$$

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

GPDOCK: highly accurate docking strategy for metalloproteins based on geometric probability, briefing in bioinformatics, 2023

Evaluation Procedure

Overall Accuracy

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision (Positive Predictive Value)

$$Precision = \frac{TP}{TP+FP}$$

Recall(Sensitivity)

$$\text{Recall (Sensitivity)} = \frac{TP}{TP+FN}$$

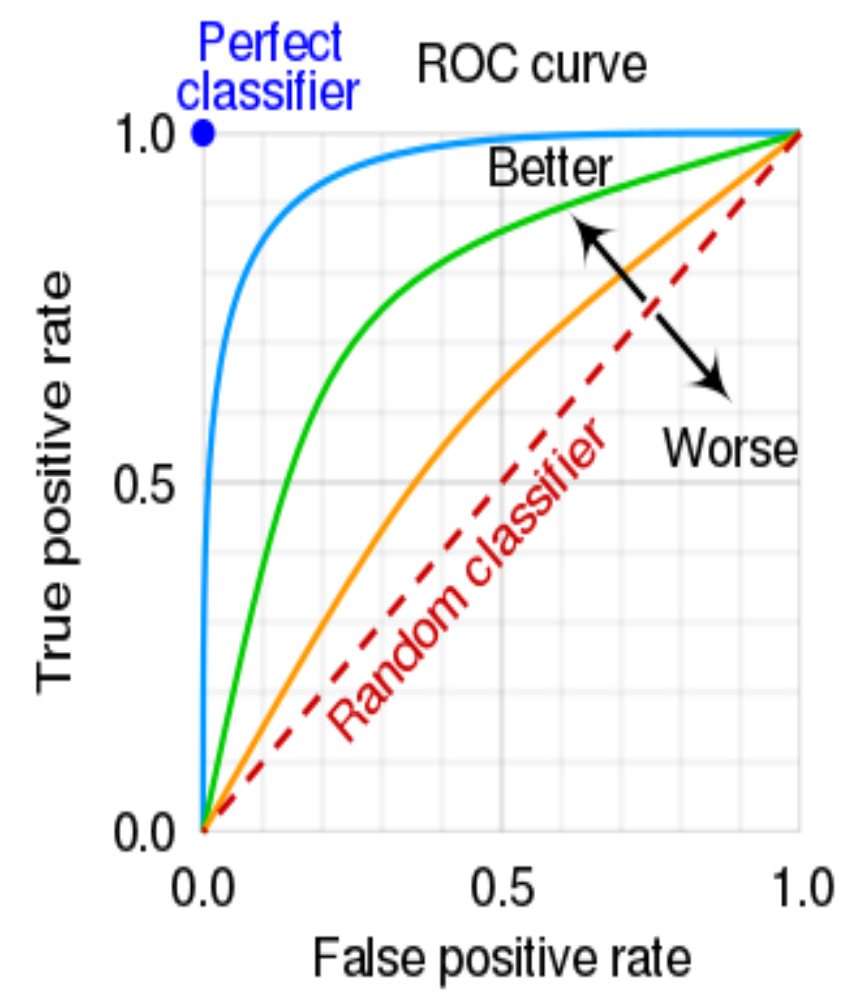
Specificity (True Negative Rate)

$$Specificity = \frac{TN}{TN+FP}$$

Evaluation Procedure

- **Receiver Operating Characteristic (ROC) Curve**
- **Area Under the ROC Curve (AUC)**
- **Recall-Precision Curve (AURPC)**
- **Matthews Correlation Coefficient (MCC)**

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$



Webserver-MAHOMES Web Server

MAHOMES II

Metal Activity Heuristic of Metalloprotein and Enzymatic Sites (MAHOMES) II - Predicts if a protein bound metal ion is enzymatic or non-enzymatic

Overview

The ability to distinguish enzyme from non-enzyme sites remains an unsolved, challenging task. We've developed MAHOMES, a machine learning based tool which classifies metals bound to proteins as enzymatic or non-enzymatic. We intend to build on the previous work to make MAHOMES II, a more stable and robust version with a web server.

System requirements

Feature calculations also require using Rosetta, FindGeo, and bluues which we run using Python 3.8 and 2.7 on RHEL 8 operating system.

Installation guide

set up virtual environment:

```
$ virtualenv --version # check for virtual enviroment
$ pip install virtualenv # download using pip if no version is found
$ virtualenv -p /usr/bin/python3 venv # create new virtual environment
$ source venv/bin/activate # switch to new env
$ pip install -r requirements.txt # add packages to environment
```

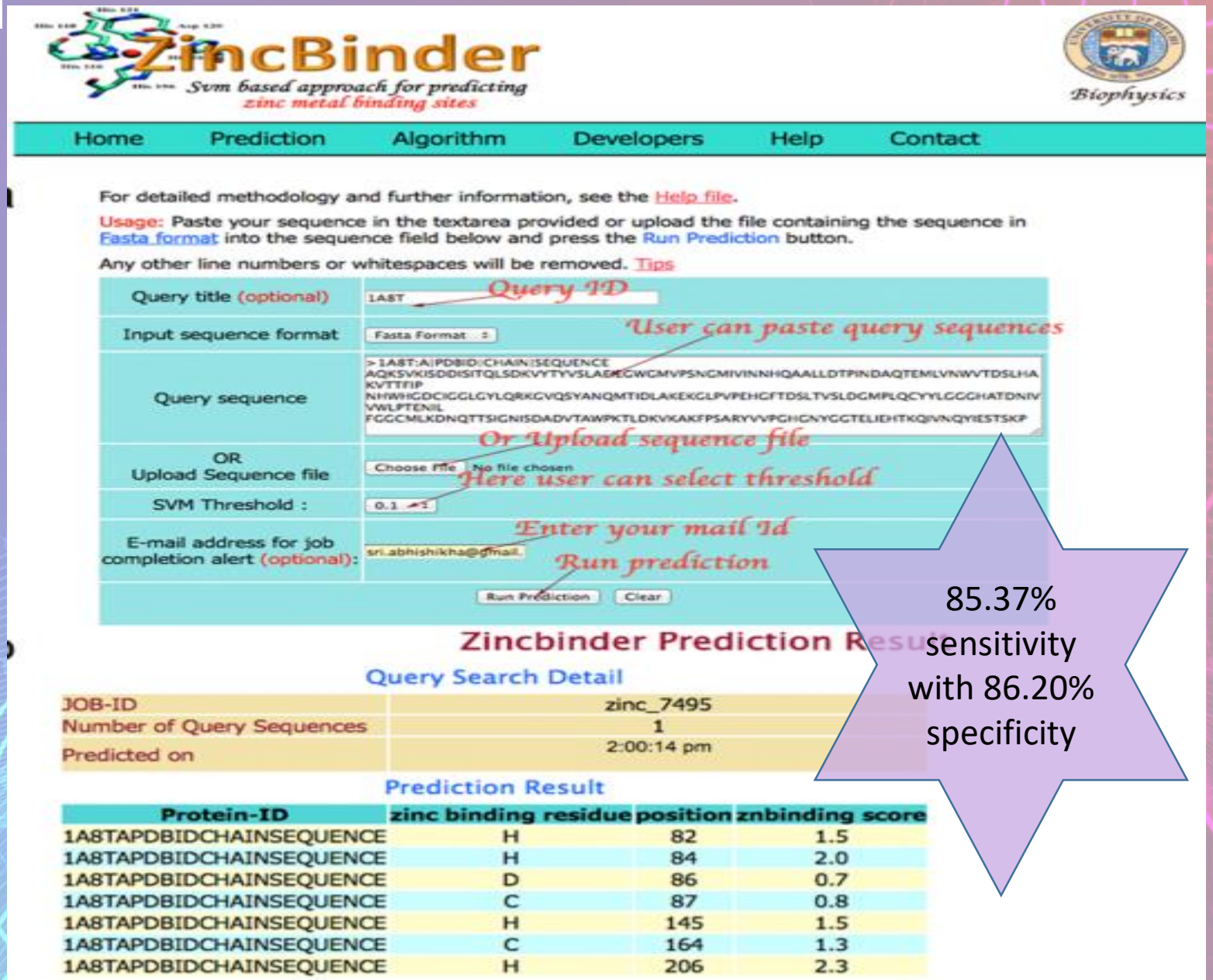
Accuracy of 90 - 97.5%

Webserver-MAHOMES Web Server

Feature group
Local protein density
solvation
Pocket void
pKa
Rosetta
Electrostatics
Pocket hydrophobicity
BLUUES SolvEnergy
Metal coordination geometry

Ponent	Description
Backend	Implemented in Python 3
Framework	Utilizes the Flask framework
Templates	HTML client-side interfaces created using Jinja
Job Metadata Storage	Stored in JSON files
Database	Metadata and scheduling information managed in an SQLite3 database
Job Execution Program	Monitors the SQLite database for new submissions, executes jobs, and sends result links via email
Web Server	Hosted on the Slusky Lab web server, operating as a virtual machine in the University of Kansas's enterprise data center

Webserver-ZincBinder Web Server



85.37% sensitivity with 86.20% specificity

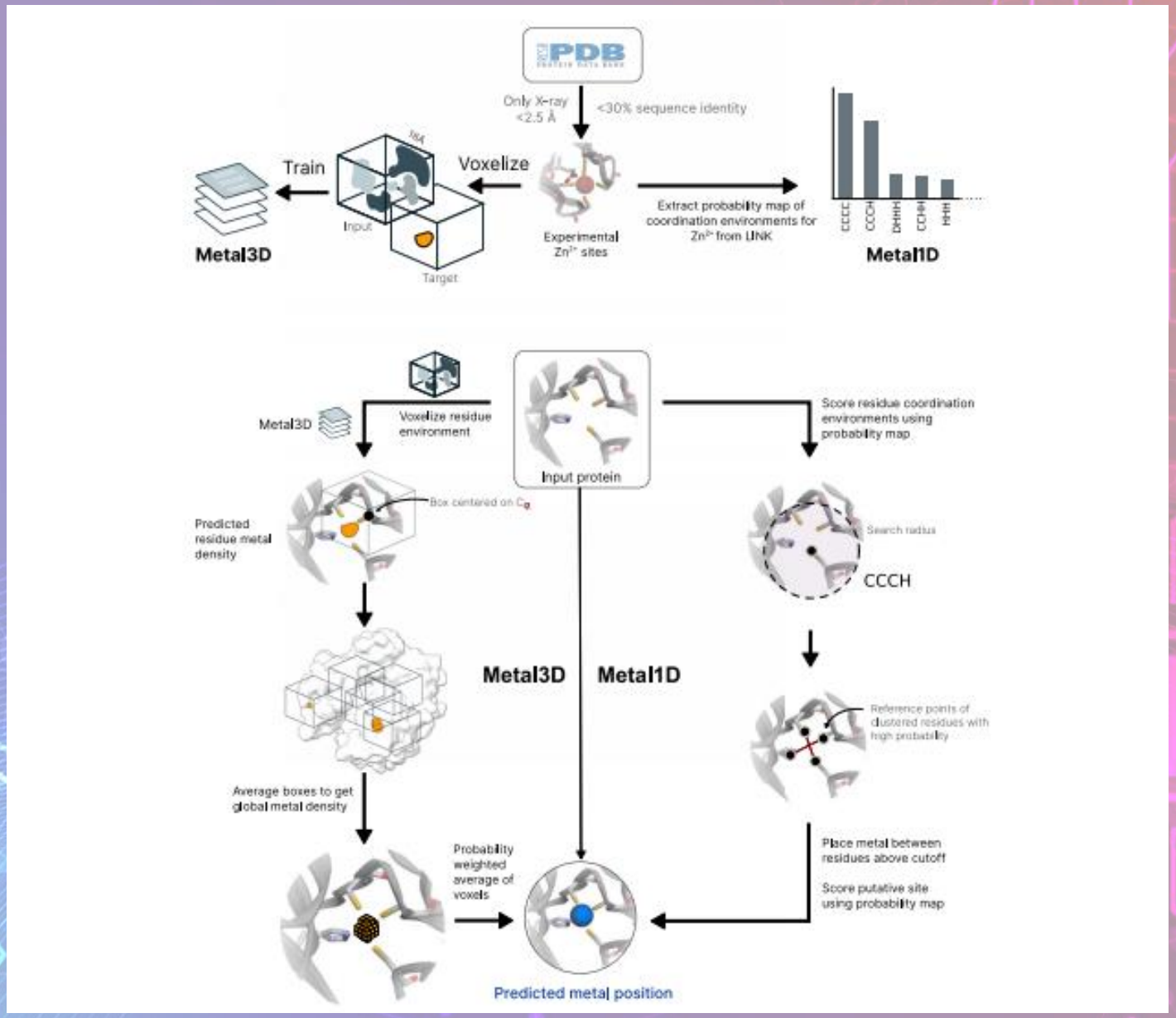
Webserver-ZincBinder Web Server

Dataset Description	Count
Total zinc-bound protein structures	1996
Total zinc-binding site IDs	3896
Protein chains obtained from PDB structures	5169
Protein chains interacting with zinc (HETEROATOM)	3924
Resolution range of protein structures (PISCES parameter)	0-2.5Å

Tools

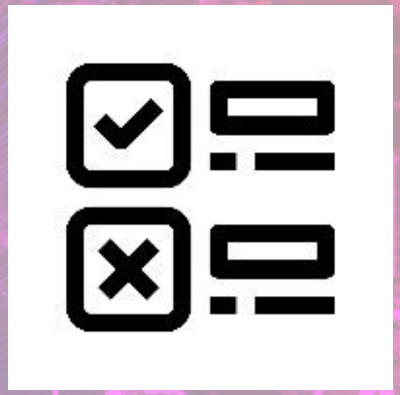
Metal Binding Site Prediction: 1D and 3D Approaches	
Metal 1D Approach	Probability Map Generation(ProbMapGenerator.ipynb), BioPandas python library
Metal 3D Approach	Voxelization Process(moleculekit Python library),

Tools



Metal3D: a general deep learning framework for accurate metal ion location prediction in proteins, nature communication, 2023

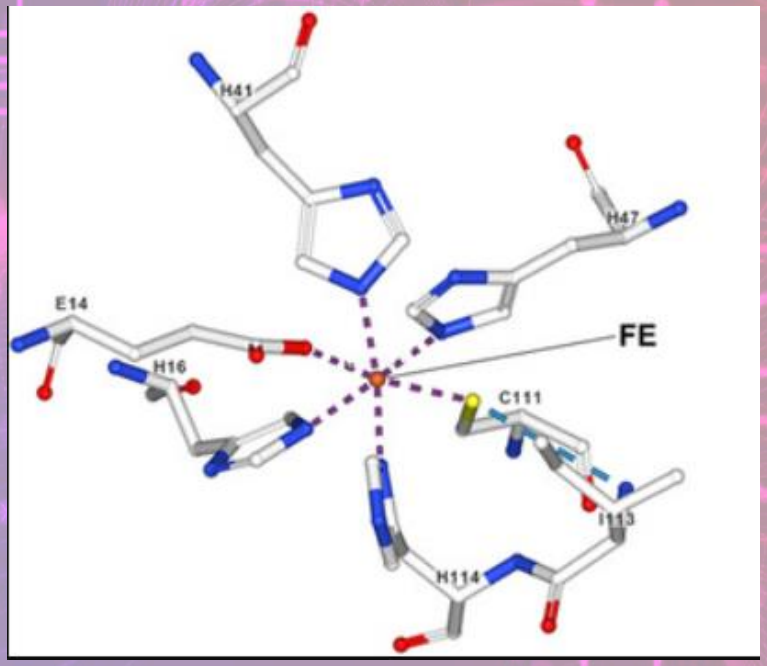
Results



Identifying Cys and His in Transition-Metal-Binding Sites Using-SVM

- Prediction of histidine in two states
- Prediction of cysteine in three states
- SVM trained to locally classify the binding state of single HIS and CYS
- Referring to metal-binding amino acids as "ligands."

Metal	CYS	HIS
Zn	46 (508/1115)	24 (374/1562)
Heme	50 (115/230)	34 (151/450)
Fe/S	63 (205/326)	3 (10/329)
Cu	33 (36/108)	32 (86/269)
Cd	62 (48/77)	32 (25/79)
Fe	13 (16/122)	18 (59/325)
Ni	4 (2/46)	16 (18/112)
Any	48 (930/1923)	25 (723/2942)

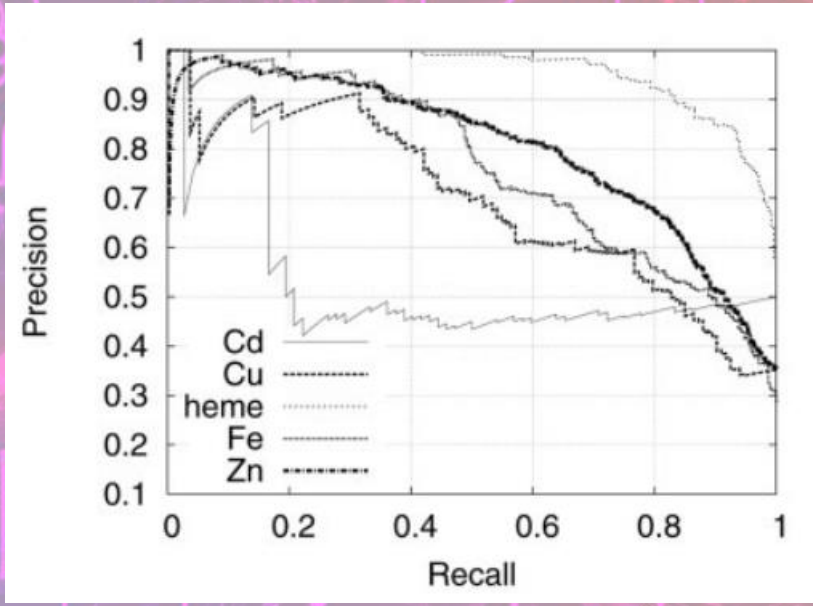


Identifying Cysteines and Histidines in Transition-Metal-Binding Sites Using Support Vector Machines and Neural NetworksPROTEINS: Structure, Function, and Bioinformatic,2006.

Prediction of Protein Metal Binding Sites Using Deep Neural Networks, Molecular information,2019

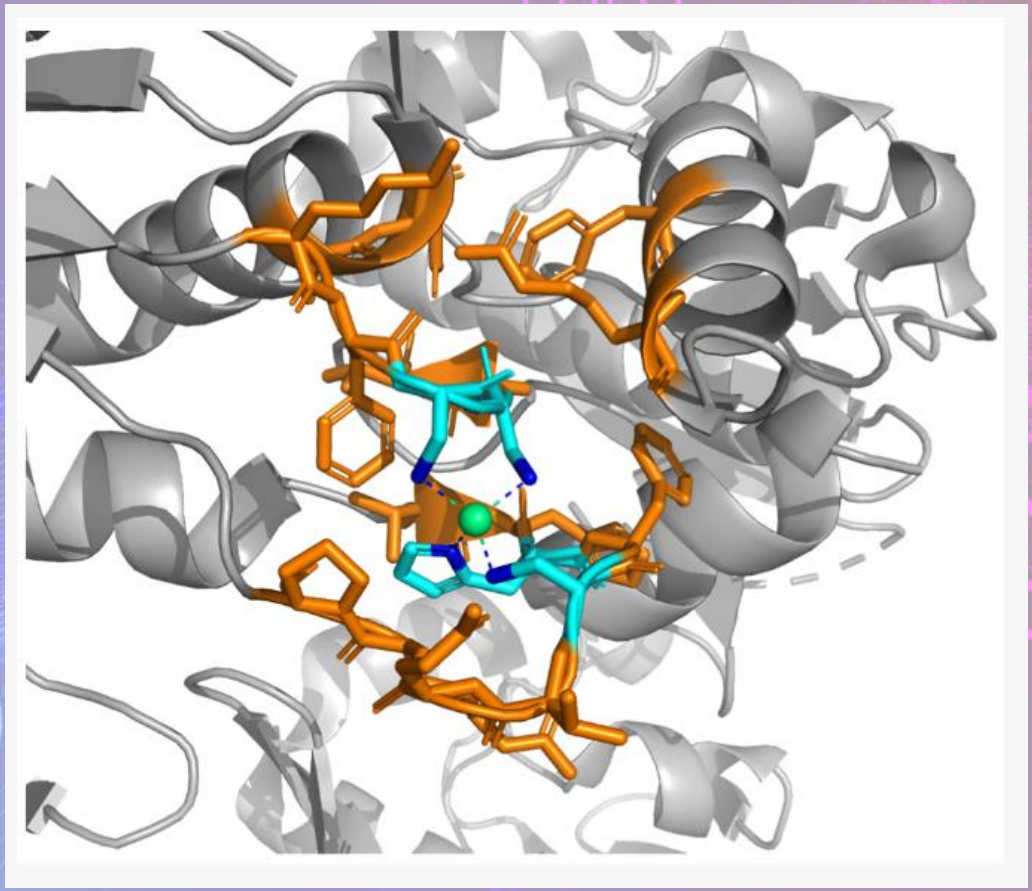
Identifying Cys and His in Transition-Metal-Binding Sites Using-SVM

Experiment Details	Results and Performance
Subset of 2982 protein chains analyzed using UniProt	Overall AUC: 0.959
SVM Tools Used	Precision (MBS): 73%, Recall (MBS): 61%
SVMLighty for binary classification (HIS)	Precision (Disulfide bridges): 86%, Recall (Disulfide bridges): 87%
bsvm for multiclass classification (CYS)	Performance loss without descriptors: 0.918 ± 0.004 AUC



Using Simplified Amino Acid Alphabets and Random Forest Model

- Challenges associated with protein 3D structure determination
- Clustering the 20 amino acids into a simplified amino acid alphabet
- Employment of a random forest algorithm



Structural Bioinformatics and Deep Learning of Metalloproteins: Recent Advances and Applications, International journal of molecular science, 2022

Using Simplified Amino Acid Alphabets and Random Forest Model

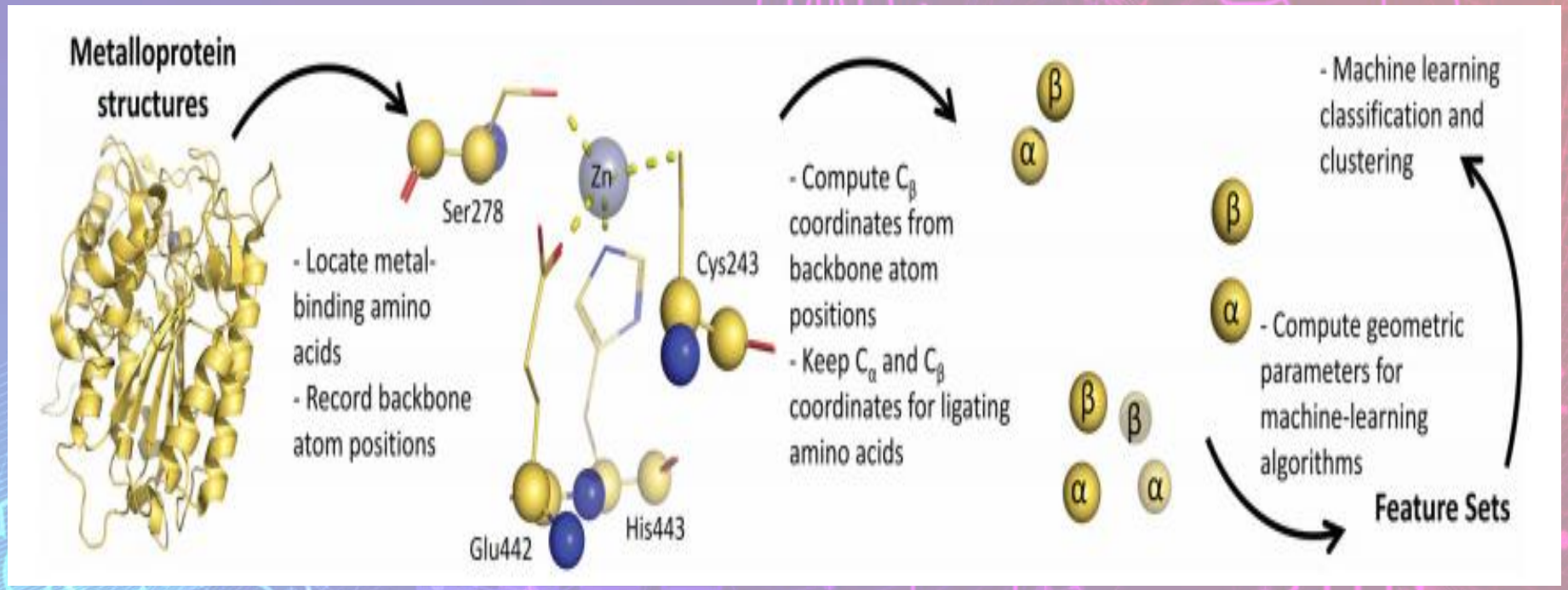
None
P
EDNQ
EDRK
PH
CILMV
AG
CFILMVW
NQSTY
CMQLEKRA

Metal Ion	Prediction Accuracy
Iron	69% ★
Copper	75%
Manganese	82%
Magnesium	80%
Nickel	90% ★
Calcium	78%
Cobalt	72%
Zinc	85%

Prediction of Metal Ion Binding Sites in Proteins from Amino Acid Sequences by Using Simplified Amino Acid Alphabets and Random Forest Model, Genomic and information, 2017

Metal binding amino acids based on backbone geometries

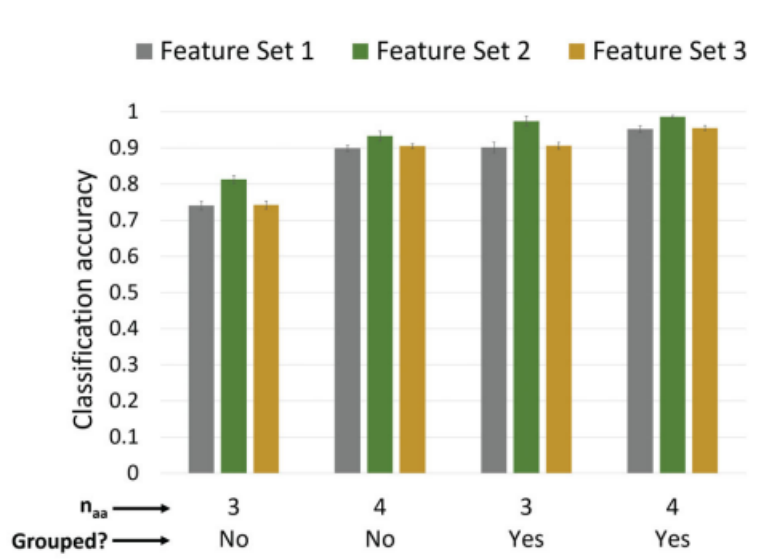
- **Decision tree machine-learning algorithm to analyze entire protein structures**



Identifying metal binding amino acids based on backbone as a tool for metalloprotein engineering, Protein ScienceY, 2021

Metal binding amino acids based on backbone geometries

Feature Set	Description	Performance
Feature Set 1	13 independent features representing the backbone geometry of coordinating amino acids	97% accuracy
Feature Set 2	Incorporates order-independent features along with the count of each type of amino acid binding to metals	
Feature Set 3	Considers all possible orderings of the coordinating amino acids, resulting in a larger dataset requiring more computational resources	

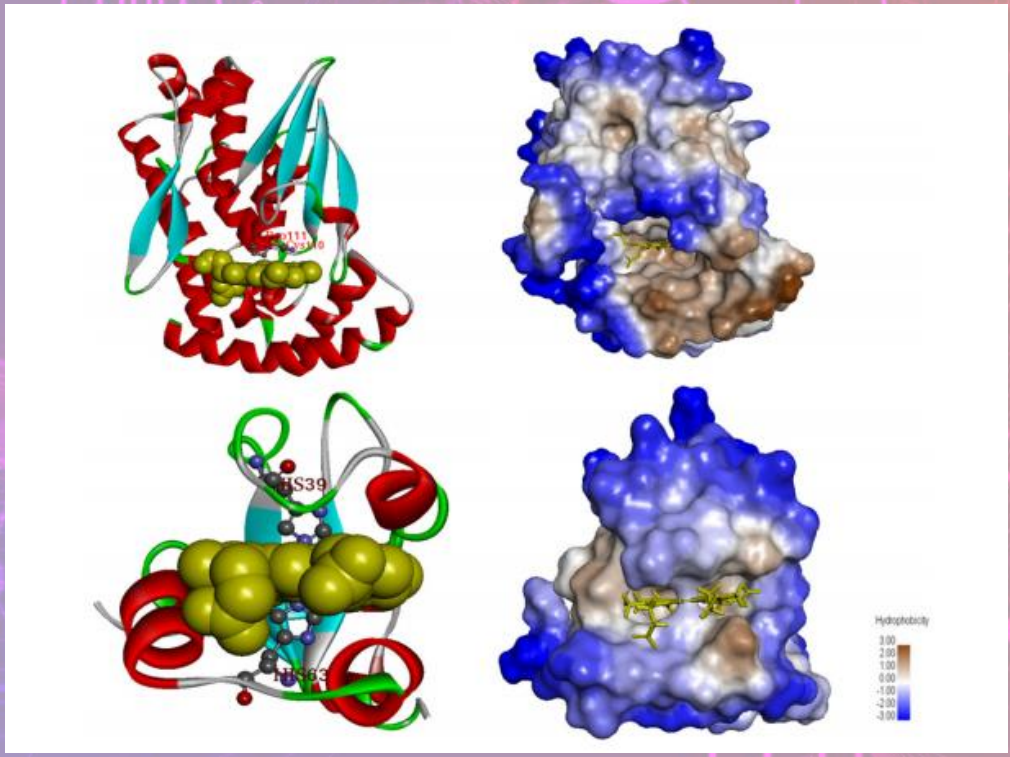


Identifying metal binding amino acids based on backbone as a tool for metalloprotein engineering, Protein ScienceY, 2021

Heme binding proteins

- **Heme binding proteins are metalloproteins containing heme ligands.**
- **Computational methods for predicting heme binding residues are limited.**

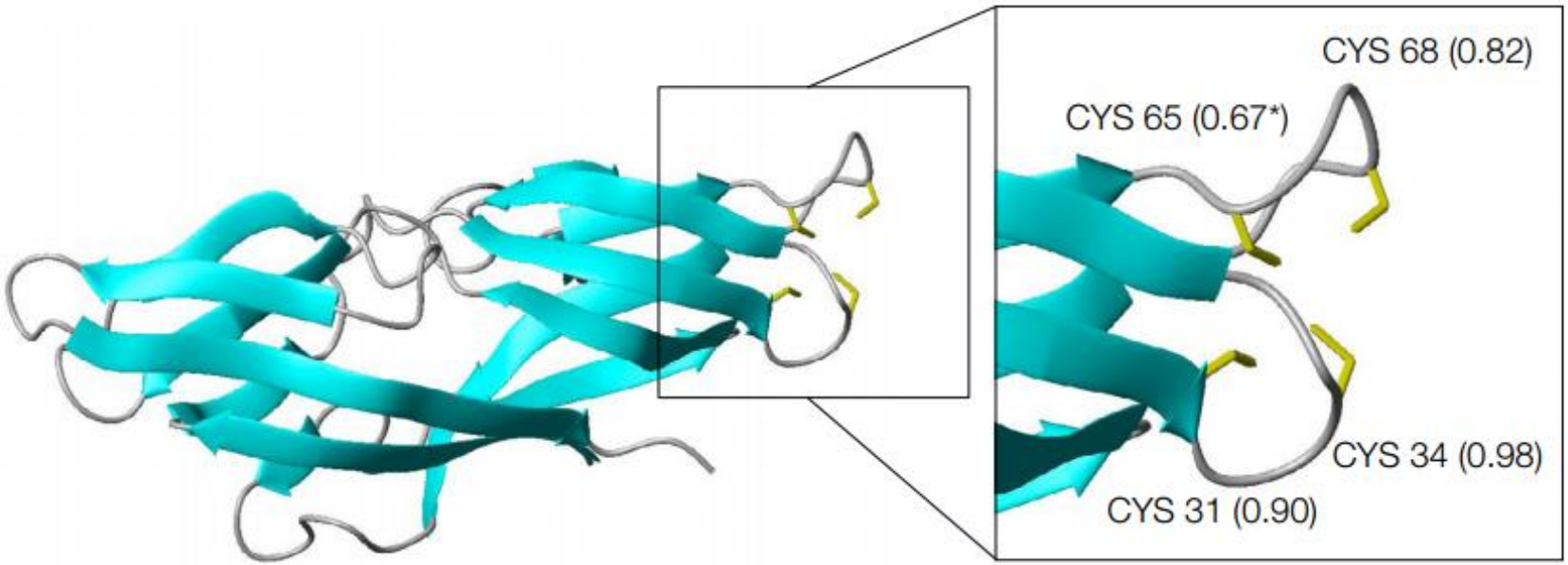
Attribute	Value/Type
Model	SVM
Dataset Size (HBPs)	747
Dataset Size (Non-HBPs)	91,414
Training Accuracy	85.90%



SCMHBP: prediction and analysis of heme binding proteins using propensity scores of dipeptides, BMS bioinformatics, 2014

Predicting zinc binding at the proteome level

- Zinc has crucial roles in catalytic and structural functions in living organisms
- A SVM approach was developed to predict zinc-binding attitudes of sequential pairs of residues

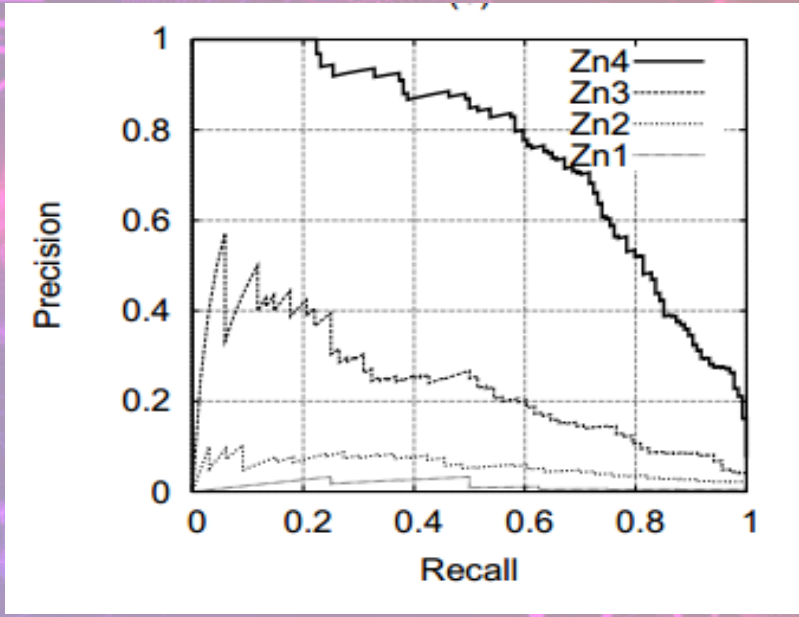


Predicting zinc binding at the proteome level, BMS bioinformatics, 2007

Predicting zinc binding at the proteome level

Procedure	Evaluation Metric
Model Selection: SVM	precision of 78% and a recall of 89%

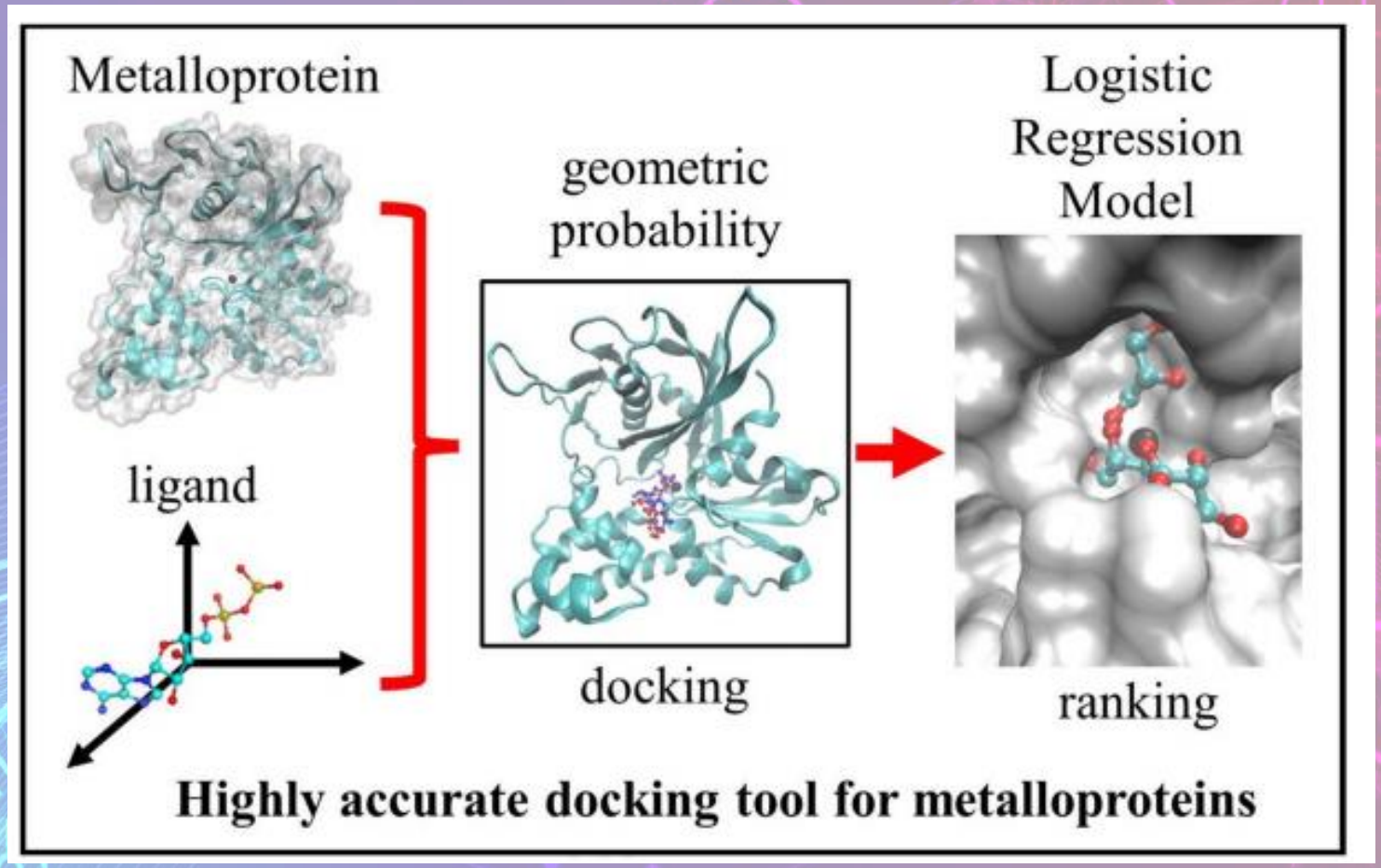
Value
Local Predictor AURPC: 0.428, AUC: 0.867 ± 0.007
Cross-validation AUC: 0.867 ± 0.007
2,833 putative zinc-binding human chains



Predicting zinc binding at the proteome level, BMS bioinformatics, 2007

GPDOCK

- **A docking method called GPDOCK (Geometric Probability Docking) is introduced, boasting unprecedented accuracy.**



GPDOCK: highly accurate docking strategy for metalloproteins based on geometric probability, briefing in bioinformatics, 2023

GPDOCK

Aspect	Description
Docking Accuracy	GPDOCK achieves 94.3% accuracy in predicting binding poses for 10 metal ions and 9360 complexes.
Docking Capability	Accurately docks metalloproteins with ligands, even with water molecules in metal ion coordination.
Dependency	Relies solely on protein and ligand structures, boosting computational efficiency.
Computational Efficiency	Employs a machine learning model for efficient scoring of binding poses.
Effectiveness	Effective and efficient for drug design and studying metalloprotein binding mechanisms.

GPDOCK

Method	Description
Datasets	Analyzed 48,184 protein structures from PDB (December 2020) with 10 metal ions.
	Coordination information obtained from crystal structure files.
Test Datasets	SM dataset: Metalloproteins with one metal ion in docking pocket.
	MM dataset: Metalloproteins with two to four metal ions in docking pocket.
	SW dataset: Metalloproteins with one metal ion and water coordination in docking pocket.

Future Directions

- **Incorporation of ligand features in articles: Many existing studies neglect to account for crucial ligand features in their analyses, hindering accurate prediction of binding sites in metalloproteins.**

- **Integration of ion coordination: Current models often overlook the coordinated fixation of ions, which is vital for understanding metalloprotein function and should be incorporated into future machine learning approaches.**

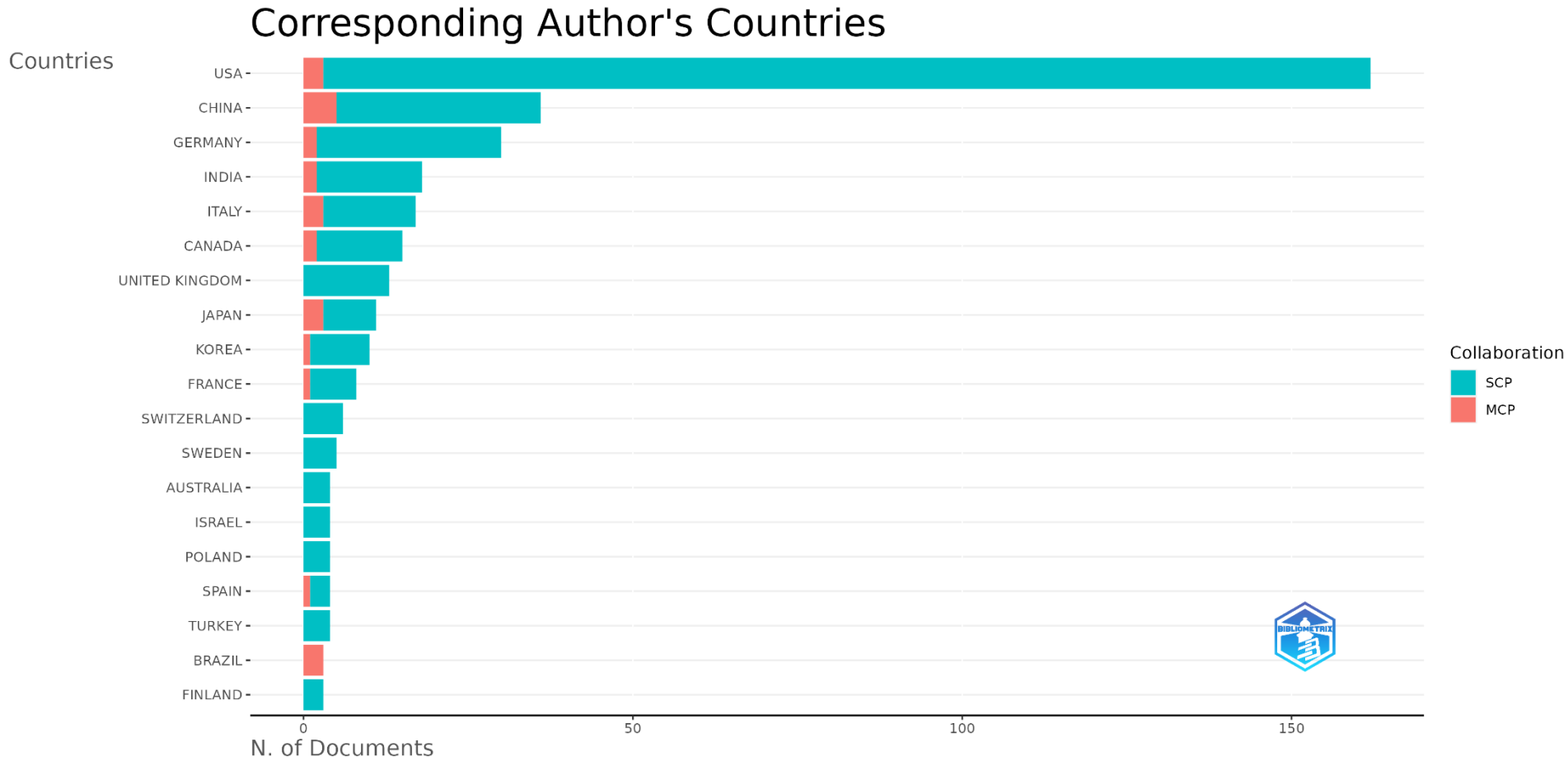
- **Machine learning-based design of new protein sequences: Future research should explore machine learning techniques to design novel protein sequences with optimized binding capabilities for specific ligands, thus advancing protein engineering efforts in the development of metalloenzymes.**

PubMed Keywords

- **Keywords Searched:**

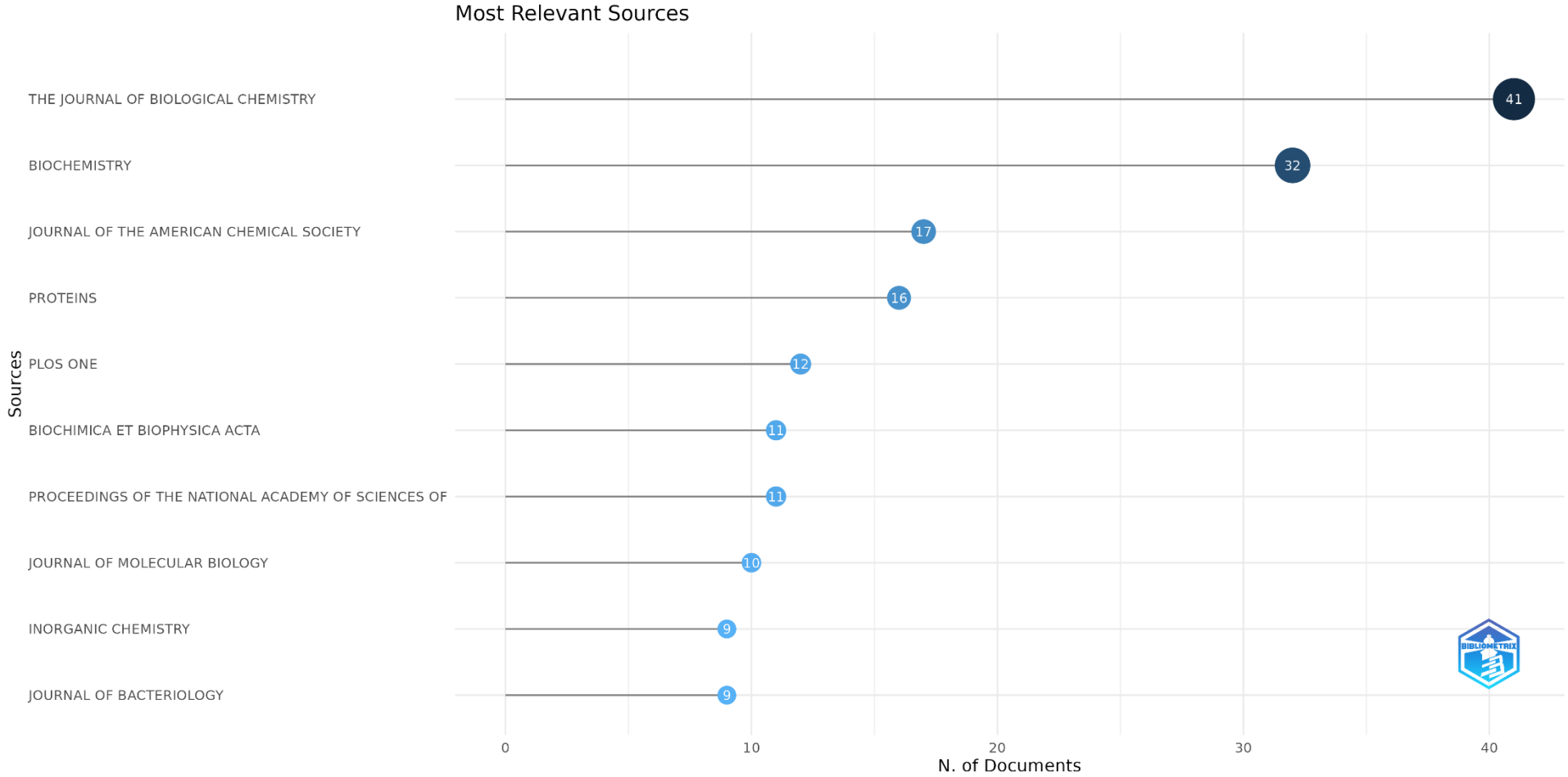
- **Machine Learning**
- **Metalloproteins**
- **Binding Sites**
- **Forecasting**
- **Machine Learning AND Metalloproteins AND (Binding Sites AND (Forecasting OR Prediction))**

Research Landscape



SCP: Single Country Publications, MCP: Multiple Country Publications

Research Landscape



Some other references

Prediction of Zinc Binding Sites in Proteins using Sequence Derived Information, Journal of Biomolecular Structure and Dynamics,2017

Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates, Protein, 2023

Metal3D: a general deep learning framework for accurate metal ion location prediction in proteins, Nature communication, 2023


Thanks for you
attention
Thank you for your
attention
Thanks for your
attention