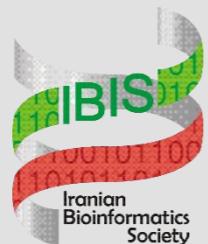
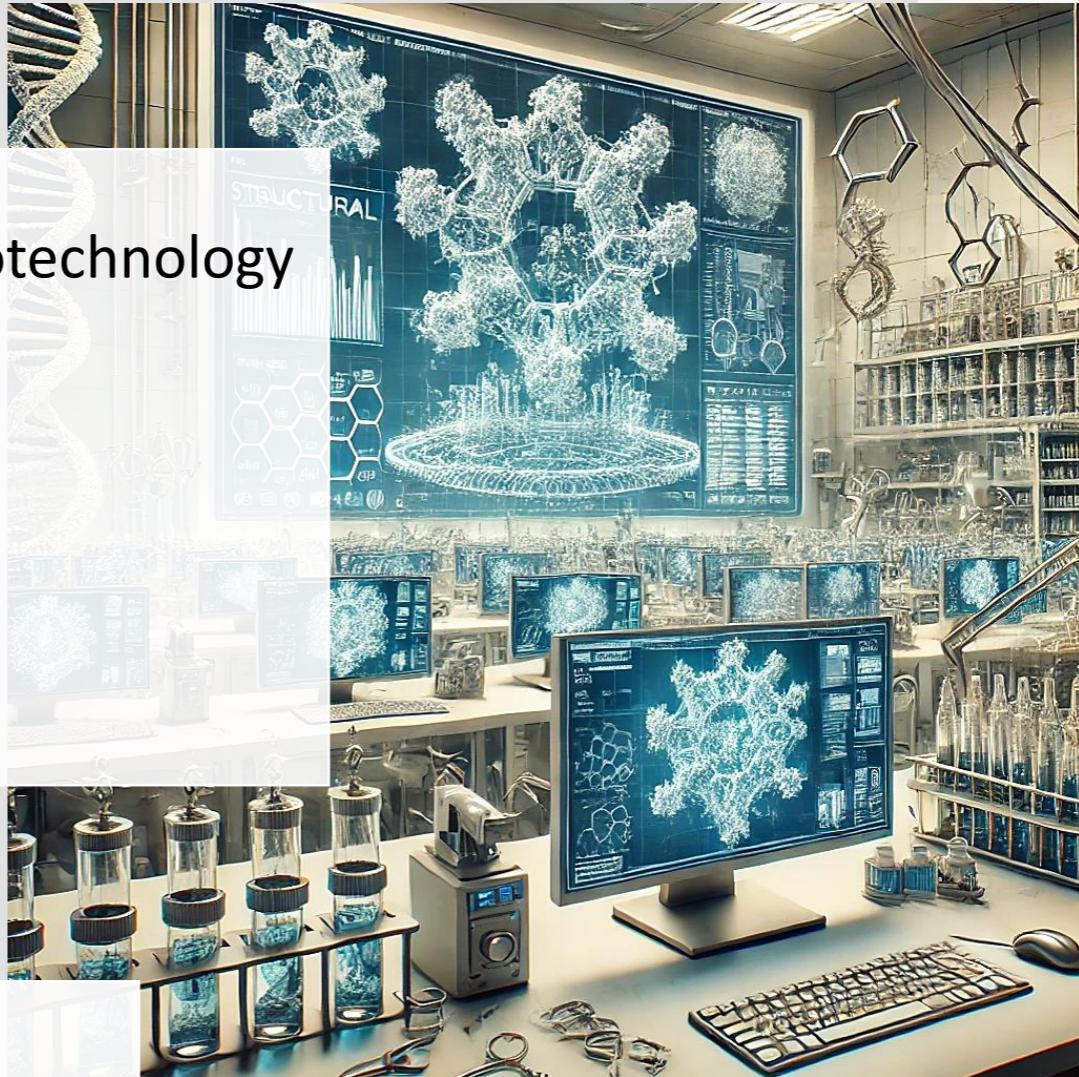


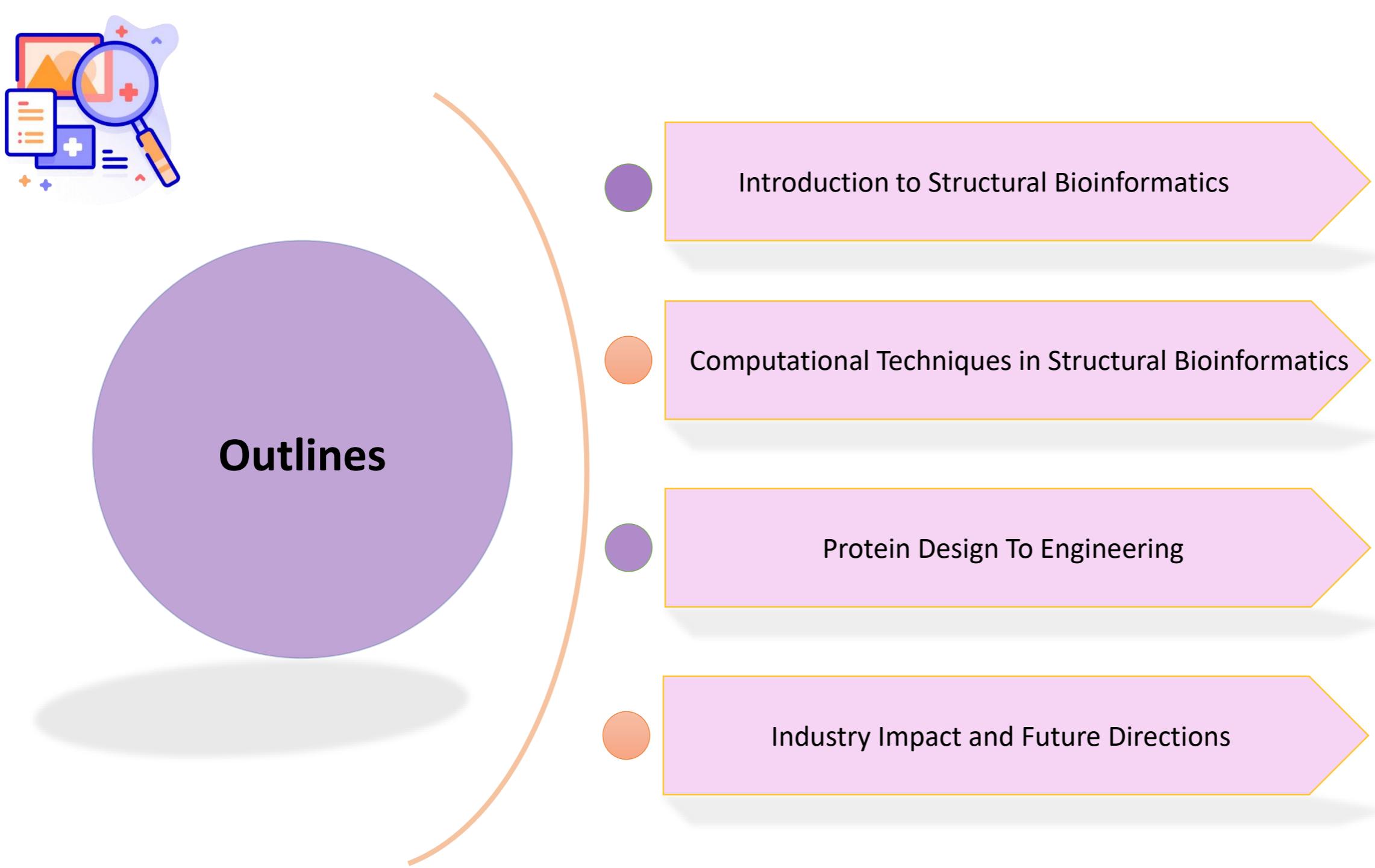
# Application of Structural Bioinformatics in the Biotechnology Industry

Supervisor: Dr. Kaveh Kavousi

Presenter: Fereshteh Noroozi

2024





# Structural Bioinformatics

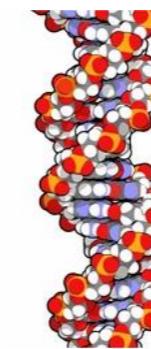
Structural bioinformatics uses computational tools to predict, model, and analyze the 3D structures of biomolecules like proteins, DNA.

... GTGCATCTGACTCCTGAGGAGAAG ...  
... CACGTAGACTGAGGACTCCTCTTC ...

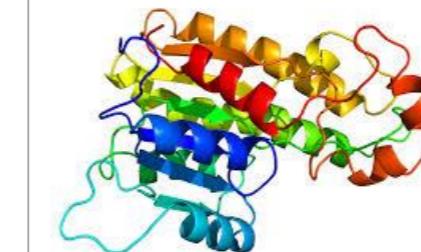


1 MEPRVVKPPQDLVVESLKSRYGLGGSCPDEYDFSNFYQSKYKRRLTSP 50  
51 GDLDIYSGDKVGSSLKYSDESKHCRTPLGSLFKHVNVNCLDDELDSFHDL 100

3D Structure Prediction



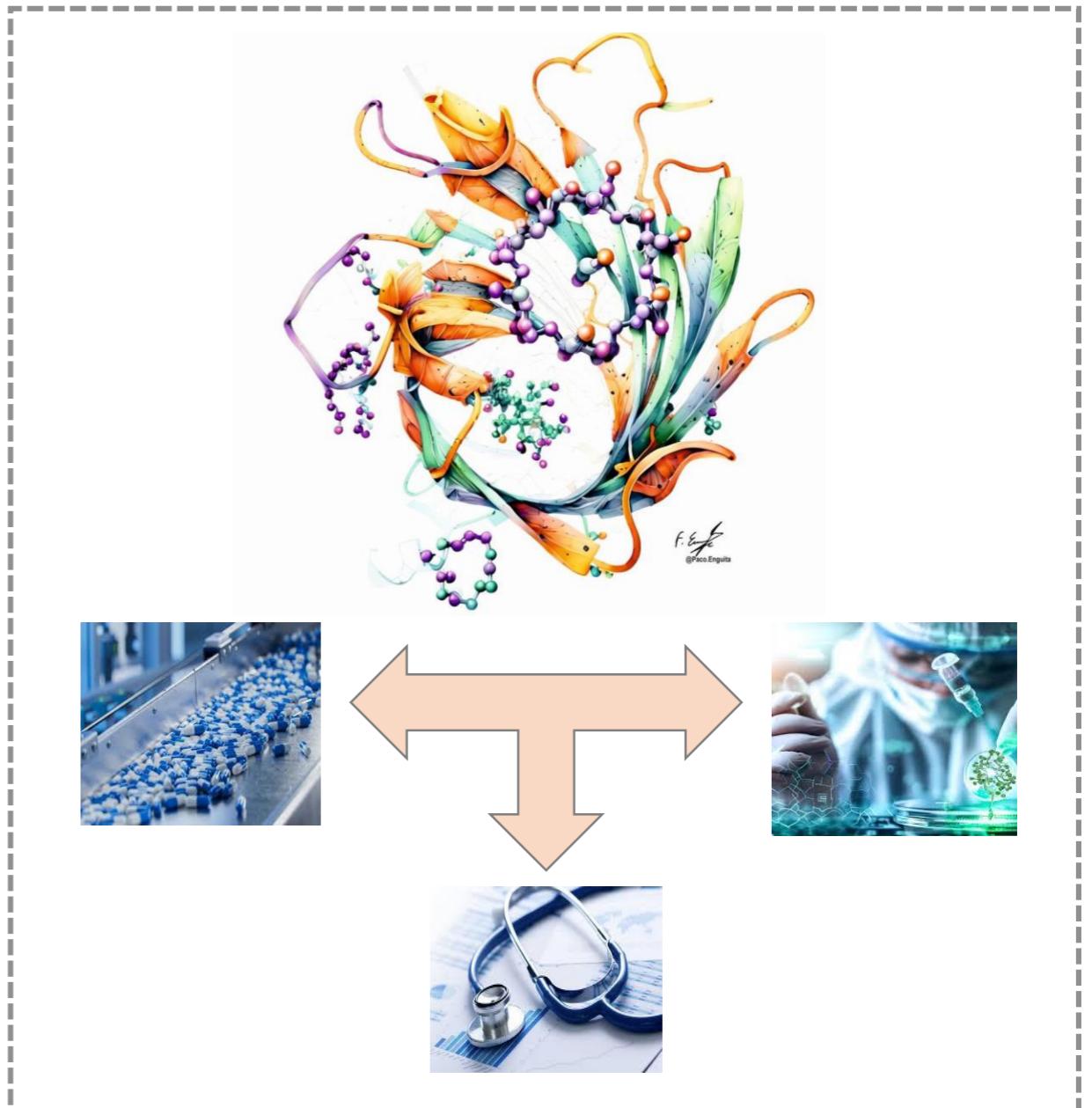
Go for further analysis



Industrial Application

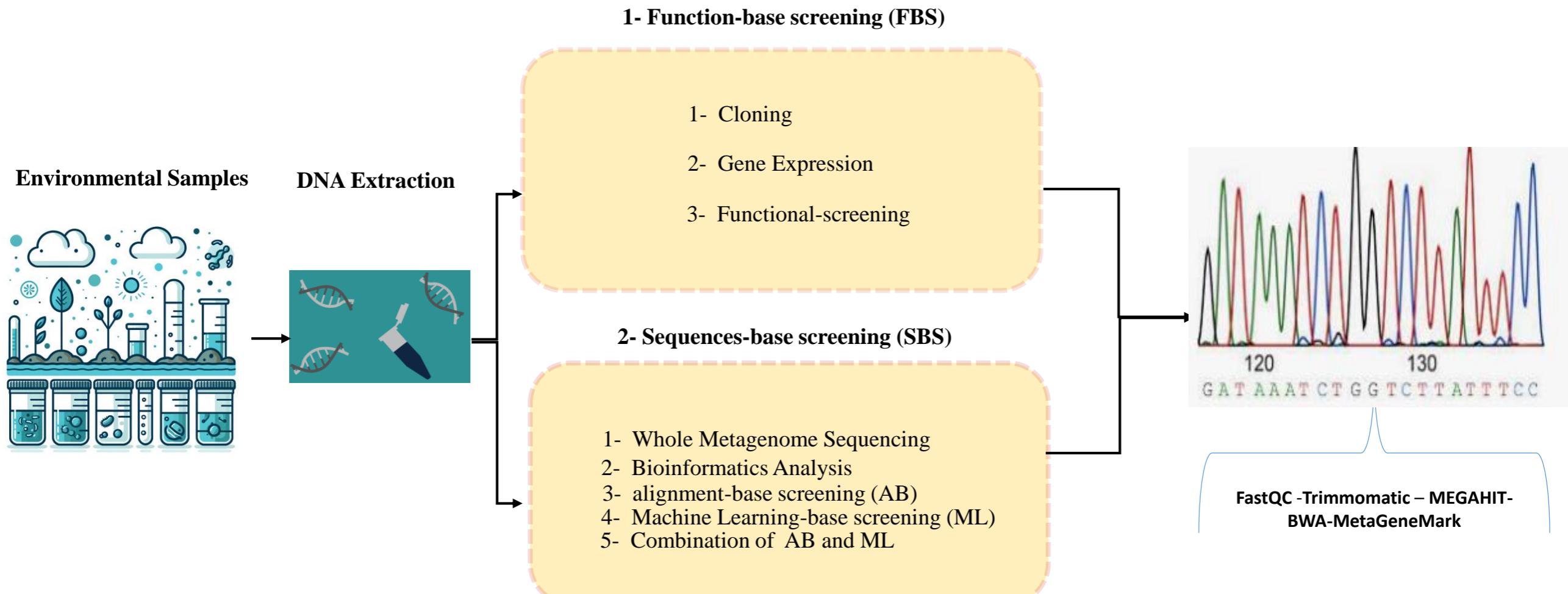
# Proteins

- ❖ Proteins are large, complex molecules made up of amino acids.
- ❖ Industries Affected by Protein Research:
- ❖ Pharmaceutical Industry
  - Wood and Paper Industries
  - Delignification and biofuel production
  - Biobleaching of Paper Pulp
  - Dye decolorization and wastewater treatment
  - Plastic degradation
- ❖ Biotechnology
- ❖ Healthcare



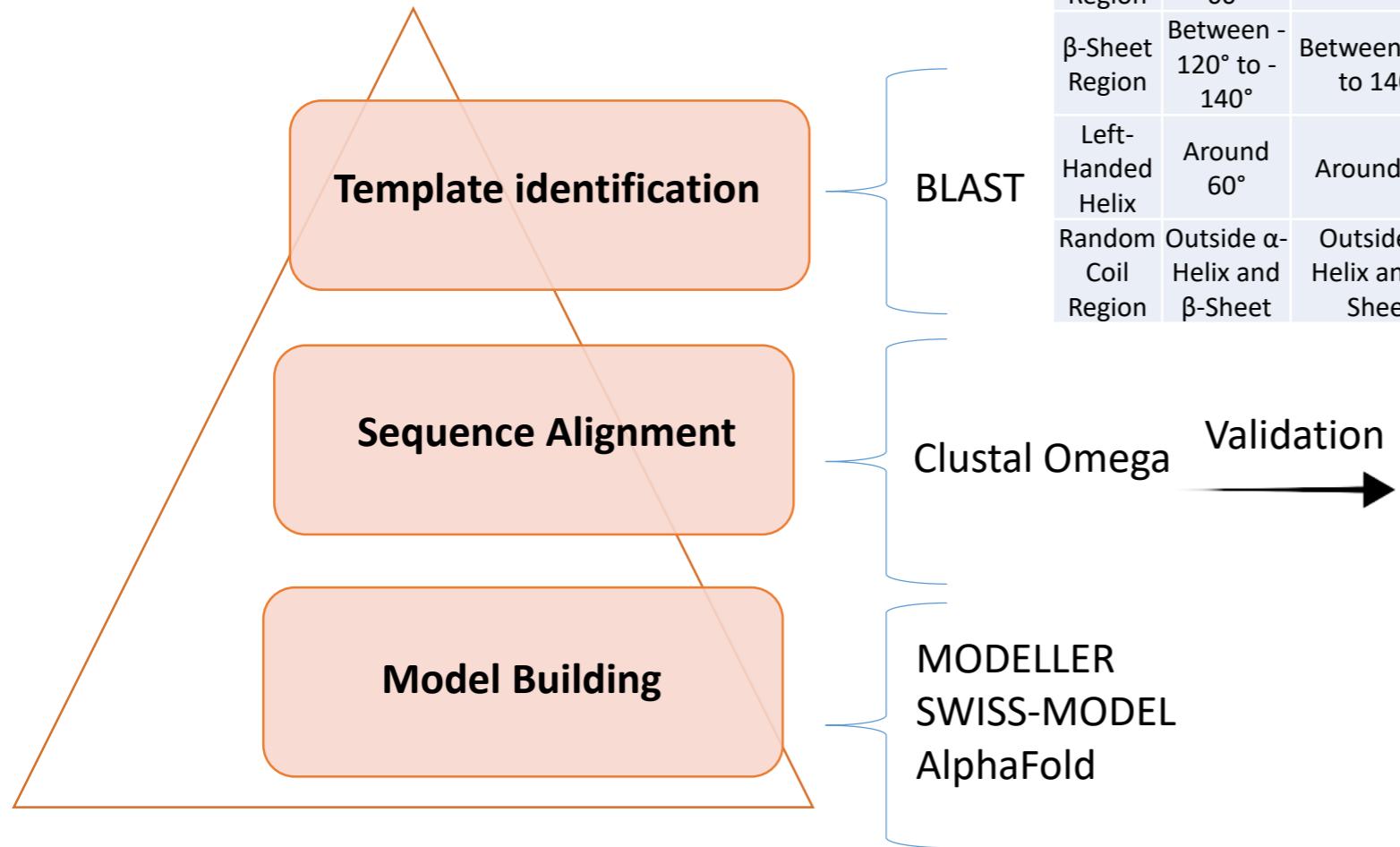
# Proteins-Sequences

- ❖ The metagenome refers to the total genetic material (DNA) present in a specific environmental sample, derived from all living organisms within that sample.

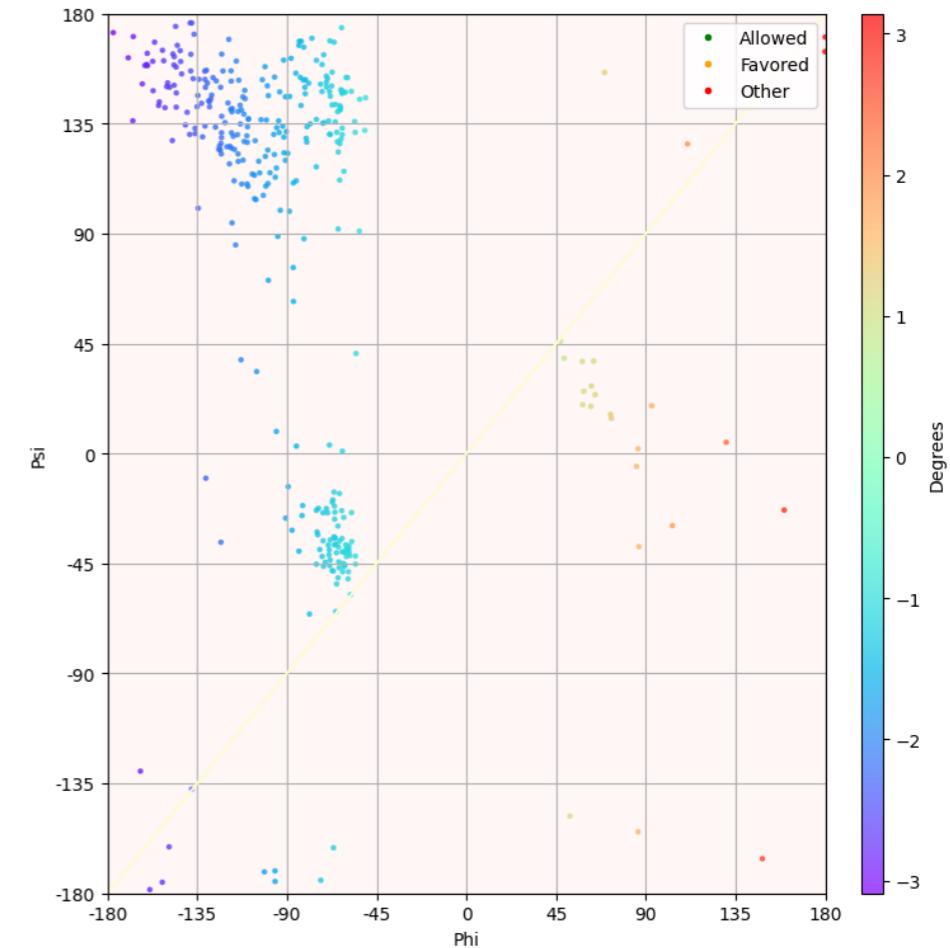


# Homology Modeling

**Homology modeling** is a computational technique used to predict the three-dimensional structure of a protein based on its similarity to known protein structures.



Region	$\phi$ Angle (Phi)	$\psi$ Angle (Psi)
$\alpha$ -Helix Region	Around -60°	Around -50°
$\beta$ -Sheet Region	Between -120° to -140°	Between 120° to 140°
Left-Handed Helix	Around 60°	Around 60°
Random Coil Region	Outside $\alpha$ -Helix and $\beta$ -Sheet	Outside $\alpha$ -Helix and $\beta$ -Sheet



# Modeller

```
>PersiLac1  
MHHHHHHHVRLRTHADTFKLKVEPGKTYLLRLINAALNDELFFSVANHTLTVVVEDAVYVKPFTVR  
TLVISPGQTNTVNLLATKPAYPGANFYMFARPYSTIRPGTFDNSTVAGILEYRNPGSLSSS  
SFDKALPIFKPMLPYFNDTNFTNFTKLRSLATKQYPAAVPQAVDRRFFFITGLGTLPC  
PKNMTCQGPNGTQFAAVNNVSVLPTTALLQSHFTGLTSVYANFPAMPLSPFNYTGT  
PPNNTHVATGTKLLALSFTNSVELVMQDTSILGIESHPLHLHGFFVVVGQFGNYDAVN  
DPAKFNLVDPVERNTVGVPAAGWVAIRFLADNPNGNRILRNPyVTNAVRTN  
  
>PersiLac2-His added  
MHHHHHHASQGRSTTALPKVCVVCIFIYLHARVLLIKQAKLTASACVRACHADTFKLKVPGKTYML  
RIINAALNDELFFSVAGHPLTIVDVAHYIKPITVETLLITPGQTNTVNLTTAKPSYPGAT  
YYMMAAPYSTAASGTFDNTTVAGILEYEDPGSPSSAGFNKNLPVLRPTLPQINDTSFVSN  
YTAKLRSLATEEYPAEVVPQEVDERRFFTGVGLGTHPCA VNGTCQGPTNDSRFAAA VNNVSF  
VLPTTALLQSHYTGMGSNGVYSSNFPAVPQSPFNTGTPPNNTVNSNGTRLVLSYGD  
LVMQGTSILGAESHPFHGHNFFVGQGFGNFPAKDPAKYNLVDPVERNTVGVPAAGW  
VAIRFRADNPNGWNPCIT  
  
>PersiLac3-His added  
MHHHHHHNSELVFVSLAGHKMTVVAADAVYTKPFETTVVLLGPQTTDVLTAVAAPGRYYLGARVY  
ASAQNVPFDNTTATAIFQYKNAAGCPPTGAGAGVGGHTGLGRPRSSGNPGRAGPAPMFFM  
LPANNNDTNTATGFSNLIRSPGPVKVPGPVQEVFTTIGFGLFCQGPFCQGPNNTRFGA  
SMNNVSFQLPNTVSLLQAHYHRIPGVFTEDFPARPPVFDYTSQNVPRALWQPVKGTRLY  
RVKYGAVVQMVFQDTGIFAAEEHPMHIGYHGYFVLTGFGNYNPRDEAKFNMVDPPSRN  
TIGVEVGGWAVVRFADNPGVWLHCHIDAHLTGGALVVEDGKTELQTTMPPPLDLP  
LCGL
```

## Enzyme Sequences

Reference Protein

## Modeller

Program for Comparative Protein  
Structure Modelling by Satisfaction  
of Spatial Restraints

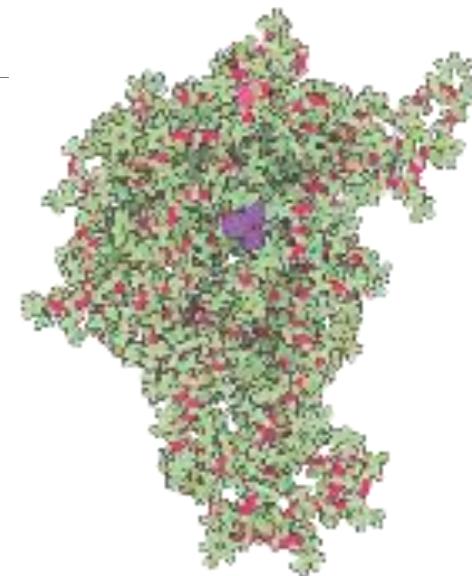


DOPE

GA341

```
from modeller import *  
from modeller.automodel import *  
  
# Create a Modeller environment  
env = Environ()  
  
# Set various parameters  
env.io.atom_files_directory = 'D:\Research-All\Laccase\modeller'  
env.io.output_directory = 'D:\Research-All\Laccase\modeller'  
env.io.hetatm = True  
env.io.water = True  
  
# Create an AutoModel object  
a = AutoModel(env, alnfile='alignment.ali',  
              knowns='6klg', sequence='lac2',  
              assess_methods=(assess.DOPE, assess.GA341))  
a.starting_model = 1  
a.ending_model = 20  
  
# Set topology and library paths  
env.libs.topology.read(file='$(LIB)/top.lib')  
env.libs.parameters.read(file='$(LIB)/par.lib')  
  
# Perform the homology modeling  
a.make()
```

Run a Python Script



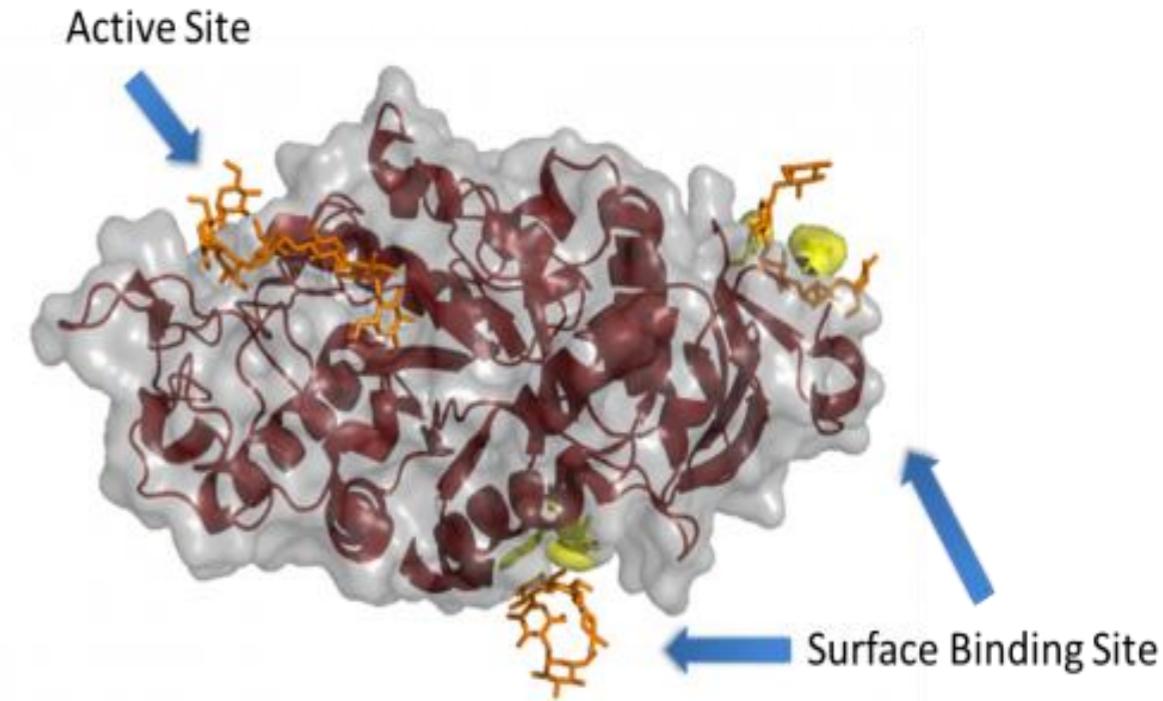
Predicted 3D structure of Enzymes

# Crucial Binding Site Significance

**What is a Binding Site?** A binding site is part of a biological molecule (such as a protein or DNA)

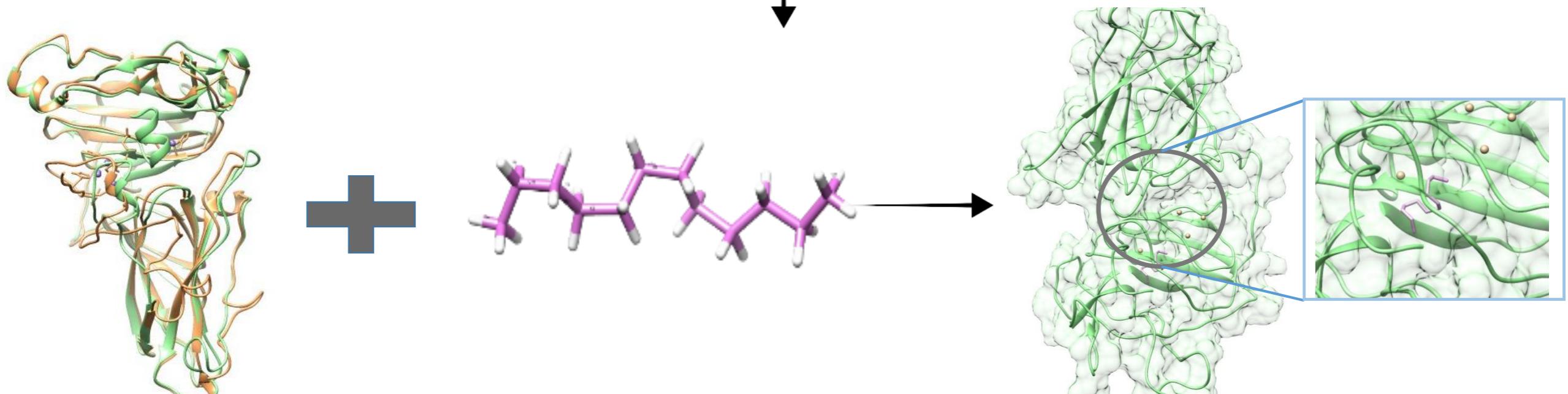
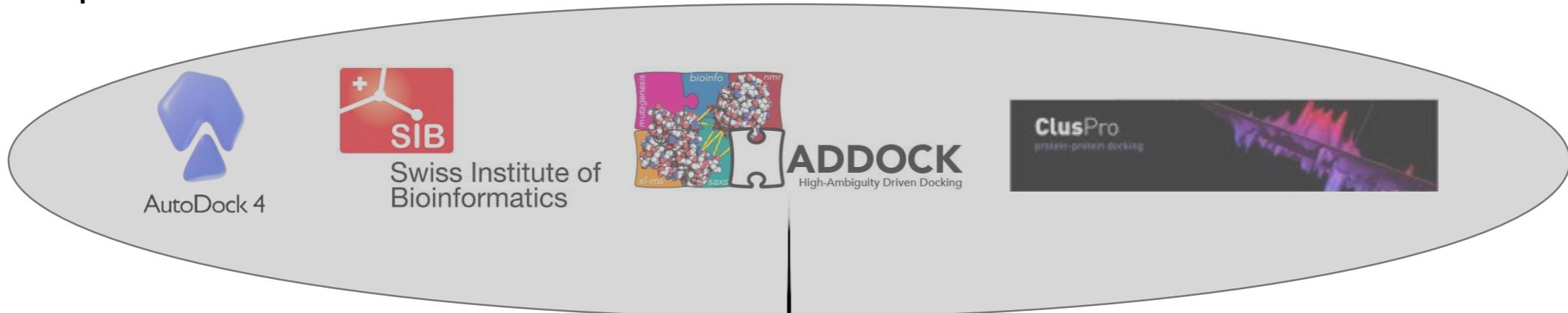
- ❖ What is a Ligand?
- ❖ What is the radius of the binding site?
- ❖ Importance of Binding Site in Biotechnology
- ❖ Diverse Methods for Binding Sites Prediction

Sequence-Based Methods  
Structure-Based Methods  
AI Approaches



# Molecular Docking

Docking is a method which predicts the preferred orientation of one molecule to a second when a ligand and a target are bound to each other to form a stable complex



Genetic Algorithm-Grid-based Search-Energy-based Scoring-Shape-based Scoring

# Computational Insights into the Selecting Mechanism of $\alpha$ -Amylase Immobilized on Cellulose Nanocrystals: Unveiling the Potential of $\alpha$ -Amylases Immobilized for Efficient Poultry Feed Hydrolysis

Seyedeh Fatemeh Sadeghian Motahar, Fereshteh Noroozi Tiyoula, Elaheh Motamed, Mehrshad Zeinalabedini, Kaveh Kavousi,\* and Shohreh Ariaeenejad\*



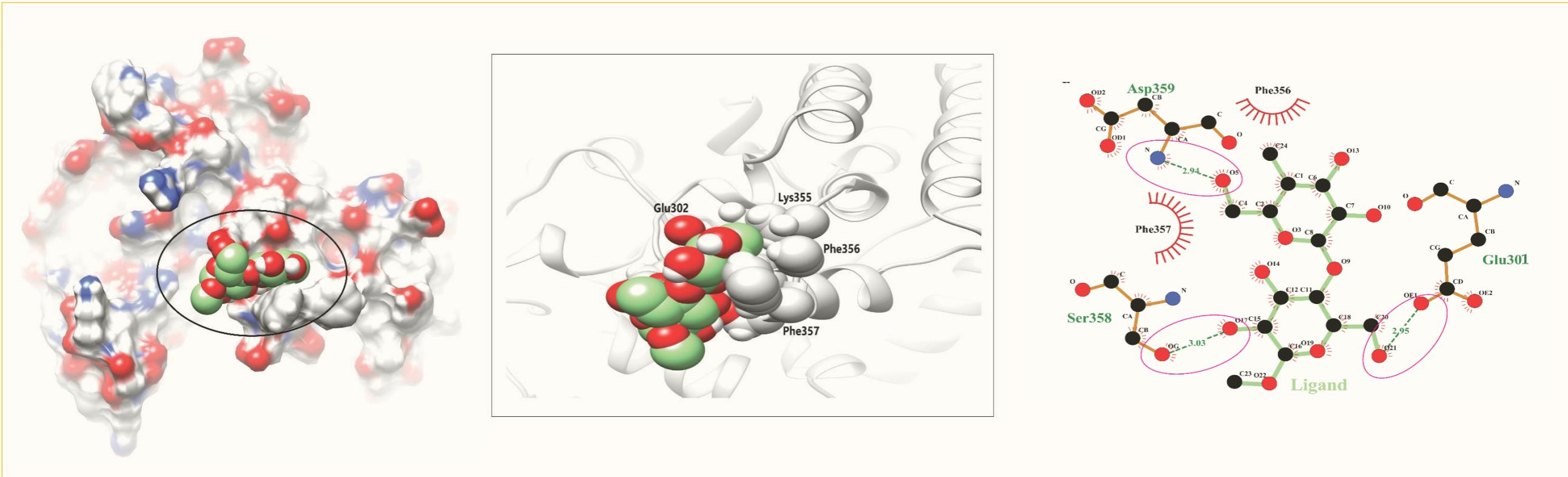
Cite This: <https://doi.org/10.1021/acs.bioconjchem.3c00304>



Read Online

- Immobilization of two  $\alpha$ -amylases (PersiAmy2 and PersiAmy3) on cellulose nanocrystals (CNCs).
- Computational methods used to compare the degradation efficiencies of the two  $\alpha$ -amylases.
- Molecular docking studies to analyze the binding affinity and interaction of  $\alpha$ -amylases with CNCs.
- Experimental *in vitro* bioconversion tests to validate the computational results.
- Examination of the effects of temperature, pH, and NaCl concentration on the activity of the immobilized enzymes.
- Investigation of the reusability of immobilized  $\alpha$ -amylases for multiple catalytic cycles.
- Assessment of the potential of immobilized  $\alpha$ -amylases for the hydrolysis of poultry feed and measuring the amount of reducing sugars produced.
- Analysis of thermodynamic parameters such as activation energy and half-life of free and immobilized enzymes.

# Case Study- Results



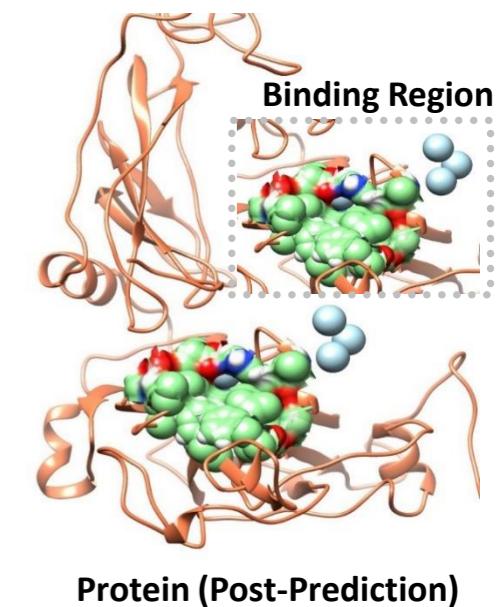
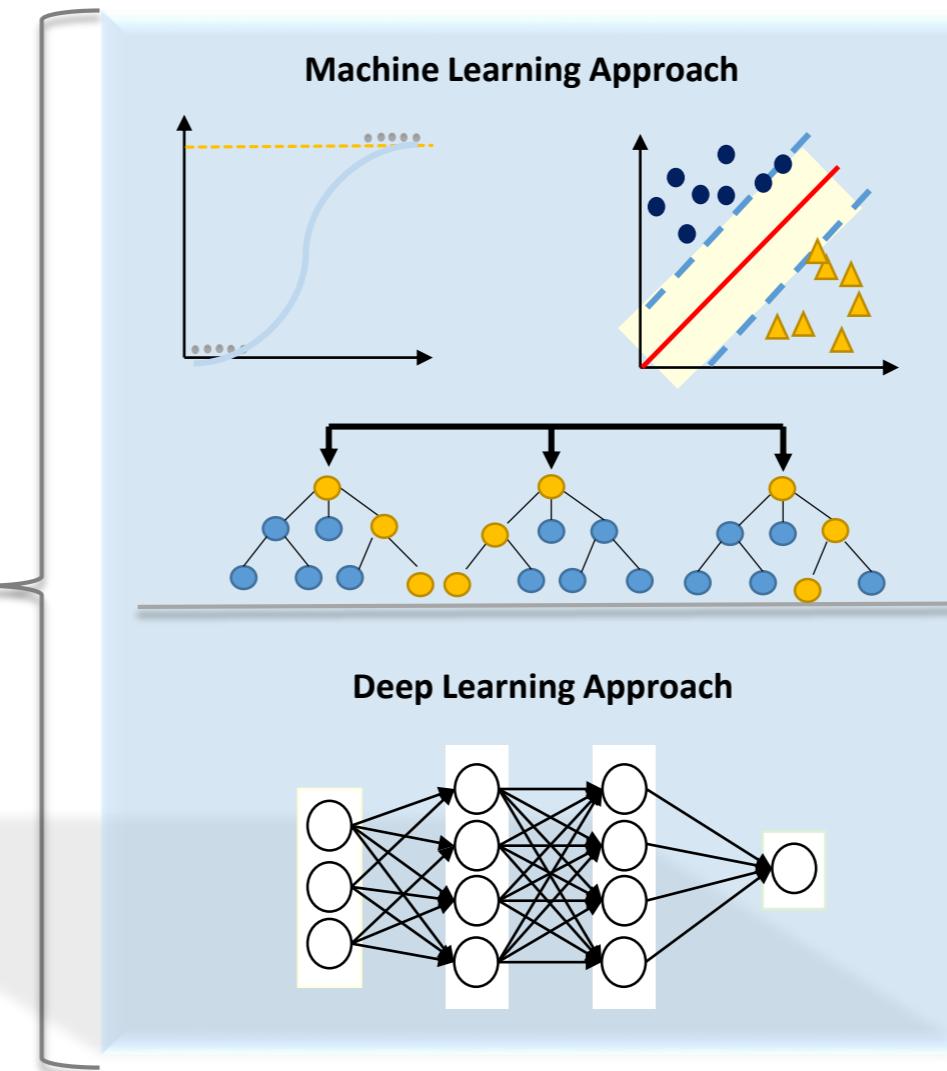
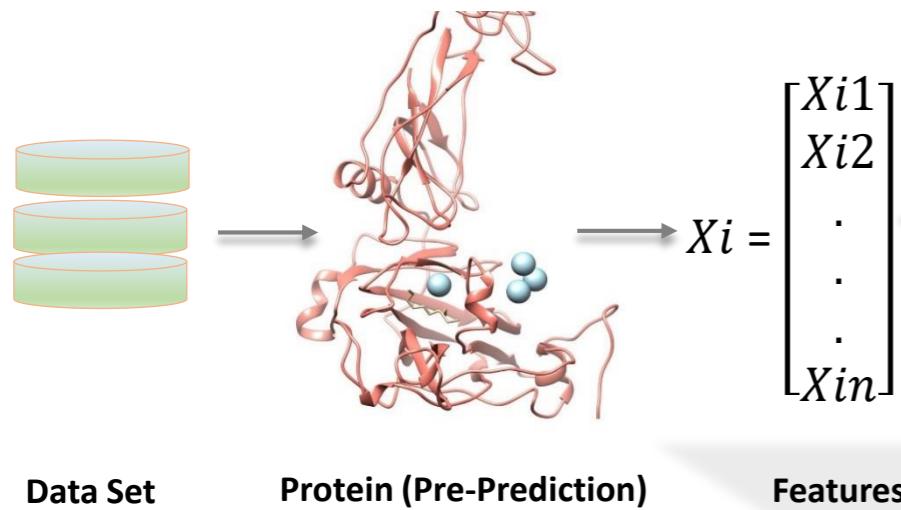
Protein and ligand contacts

→ Most important van der Waals interactions

→ Hydrogen bonds

# Machine Learning

- This process involves using machine learning algorithms to identify active and interactable regions in proteins that can bind to ligands.



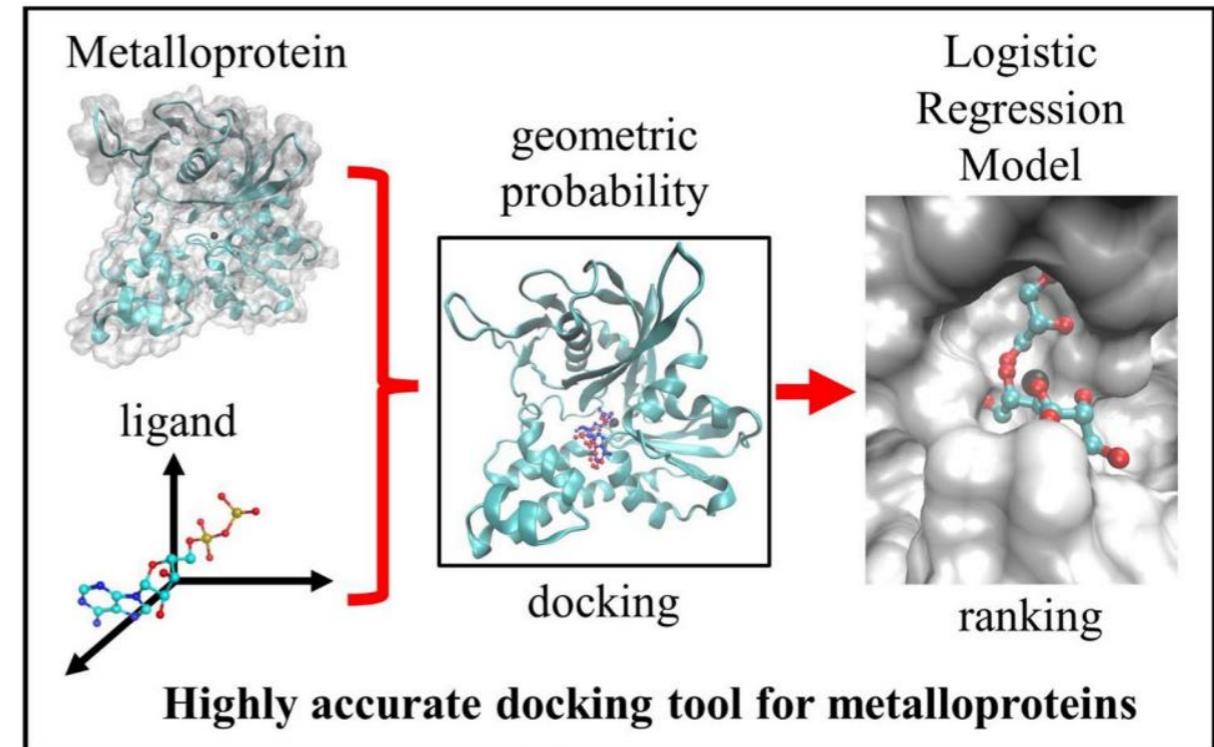
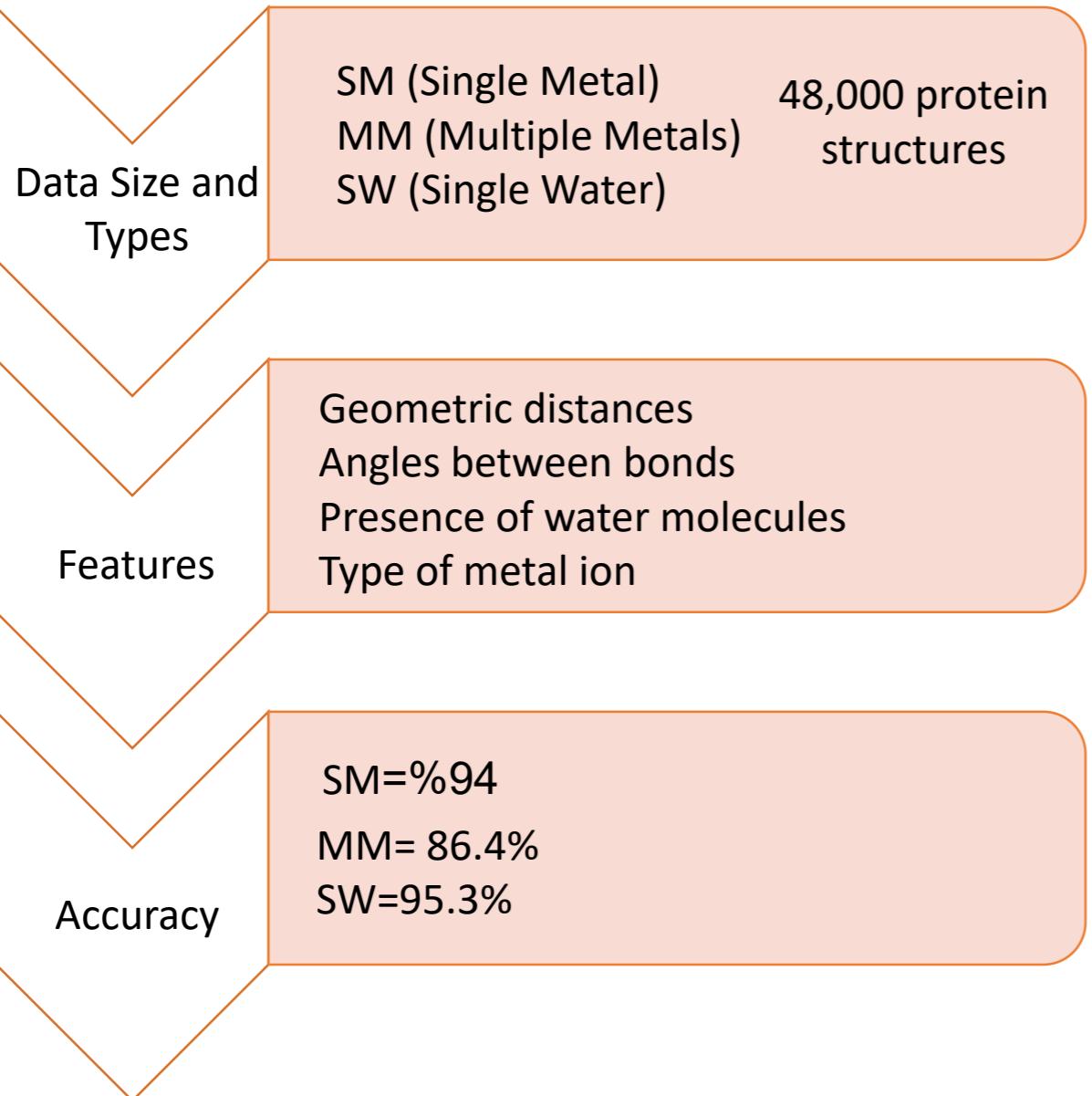
# GPDOCK: highly accurate docking strategy for metalloproteins based on geometric probability

Kai Wang 

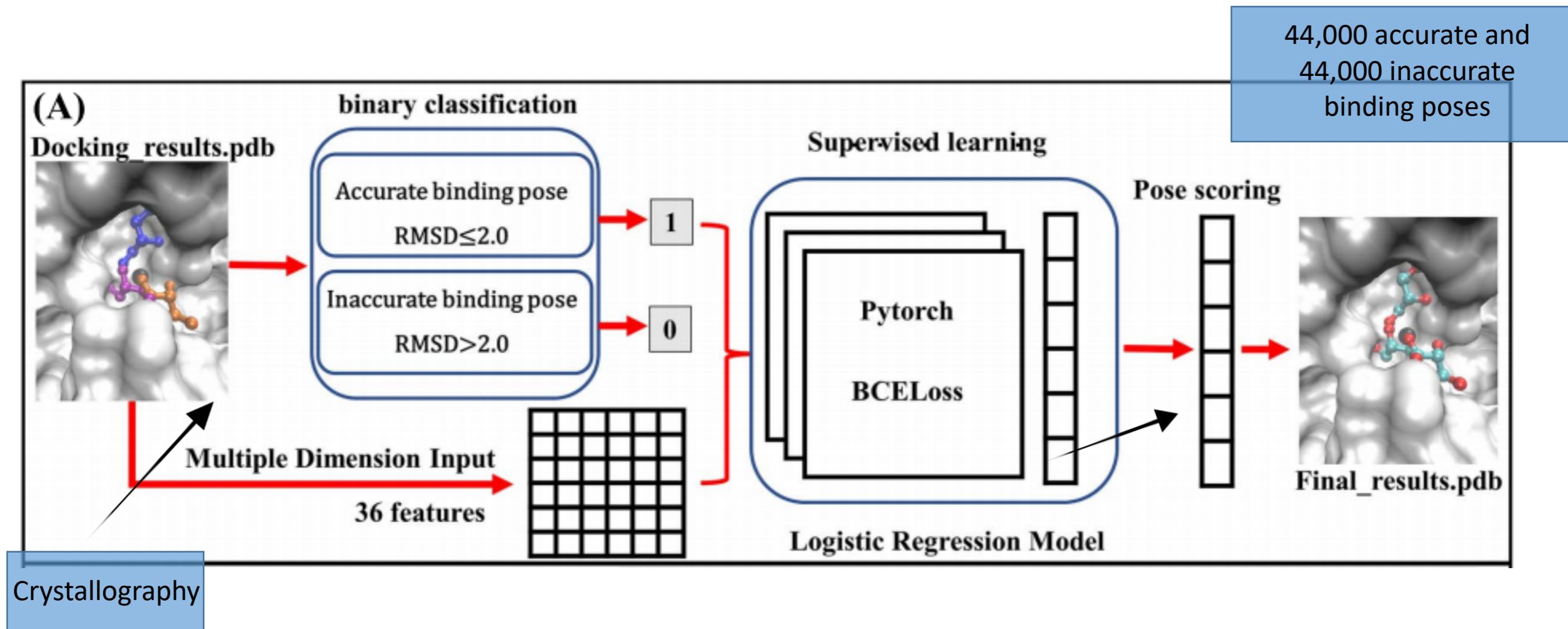
Corresponding author. K. Wang, E-mail: wangkai@zhku.edu.cn

The primary goal of this study was to develop and evaluate a docking method called **GPDOCK**, specifically designed for **metalloproteins**. A key aspect of this method is the integration of **machine learning** to enhance the process of **ranking binding poses**. By using **logistic regression**, the method automatically ranks docking results and improves the speed and accuracy of predictions, compared to traditional force-field scoring methods. The ultimate objective was to create a method that could predict ligand binding poses in metalloproteins with high accuracy and low computational resources, utilizing **machine learning** to optimize the results.

# Case Study



# Case Study



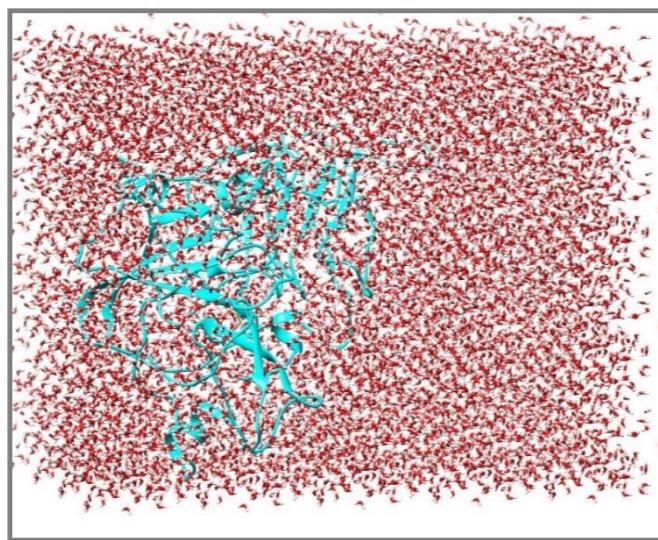
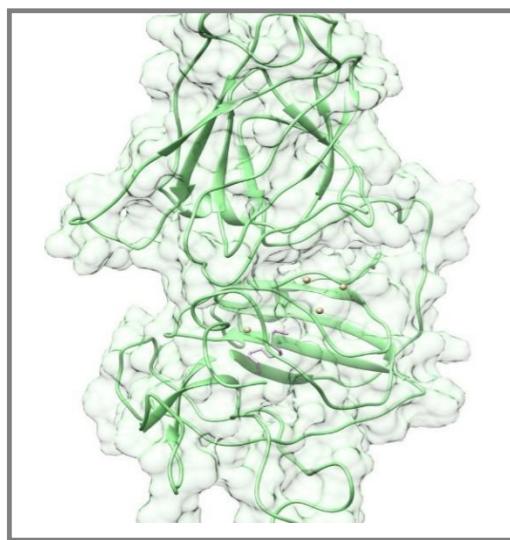
# Molecular Dynamic Simulation

- ❖ MD is a computer simulation method for analyzing the physical movements of atoms and molecules.

**GROMACS**  
FAST. FLEXIBLE. FREE.



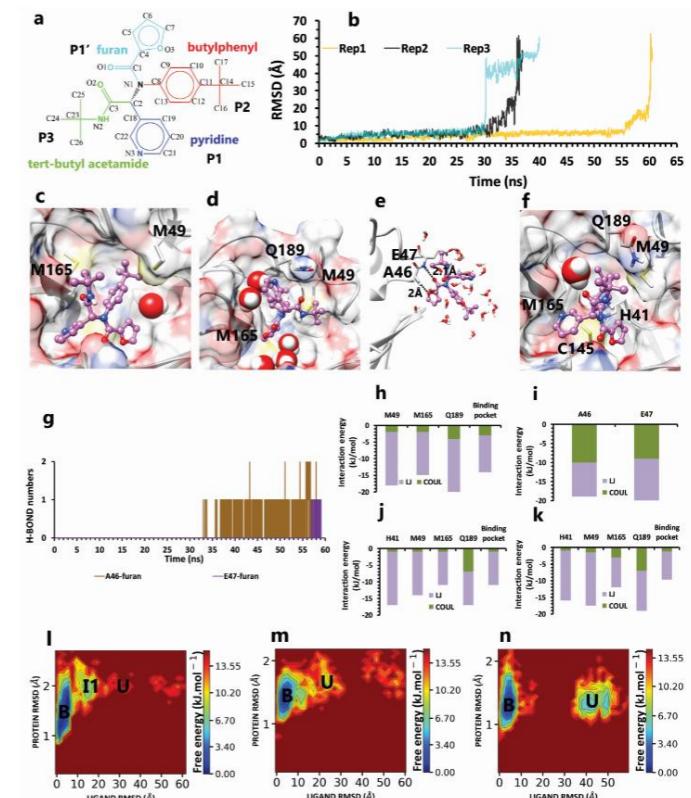
Method	Purpose
LINCS Algorithm	Maintain constant bond lengths during simulation.
Particle Mesh Ewald (PME)	For accurate calculation of long-range electrostatic interactions in large systems.
V-rescale Thermostat	To maintain system temperature at 298.15 K
Parrinello-Rahman Barostat	To control pressure at 1 bar and temperature at 310 K after temperature equilibration.

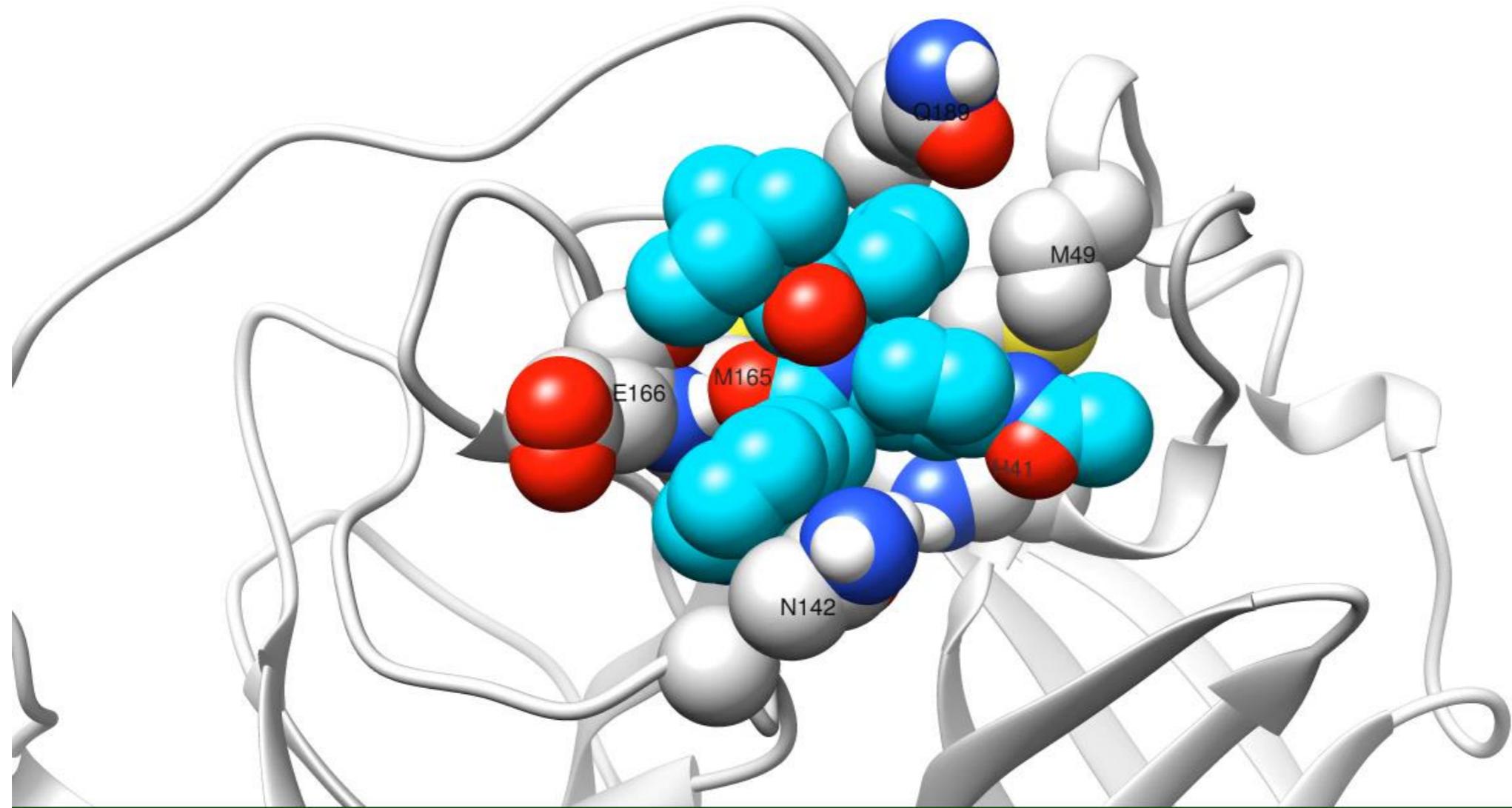


Protein-Ligand Complex

MD Simulation

Trajectory Analysis





A detailed 3D ribbon diagram of a protein structure, possibly a ribozyme, showing its secondary structure and a bound ligand. The protein backbone is depicted as a grey ribbon. A cluster of cyan, red, and blue spheres represents the bound ligand, which looks like a nucleic acid molecule. Several amino acid residues are labeled: E166, M165, N142, M41, M49, and Q189. These labels are positioned near their corresponding residues within the protein's structure.

# Exploring the Efficiency of Metagenomic Laccases in Degrading Low-Density Polyethylene (LDPE): A Comprehensive Study of Experimental and Computational Dynamics

## Key Data:

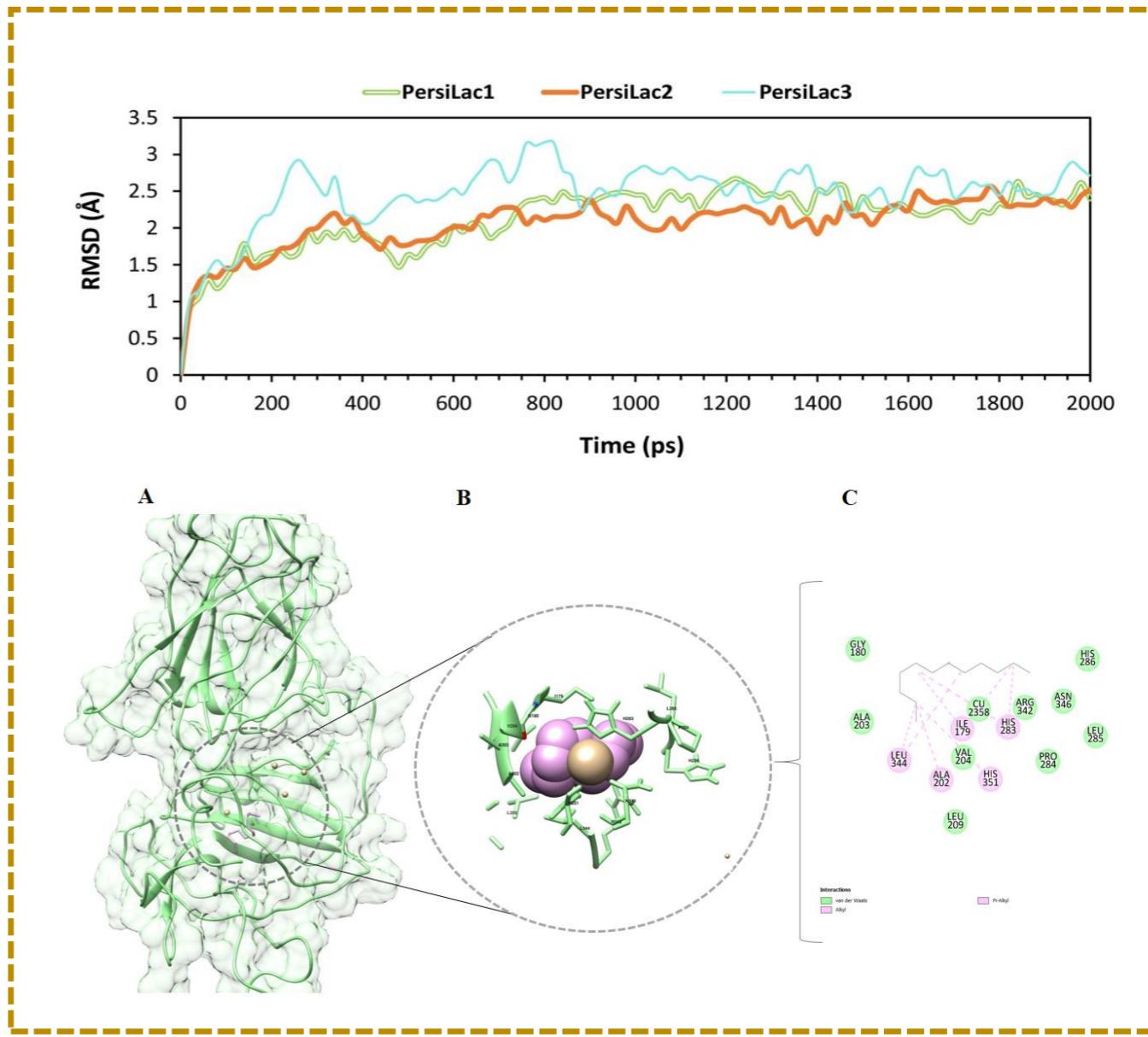
Laccases: PersiLac1, PersiLac2, PersiLac3

Methods Used: SEM, FTIR, AFM, Homology modeling, Molecular docking, MD simulations

Computational Tools: MODELLER, AutoDock, GROMACS

Experimental Assays: LDPE films, Enzyme activity assay

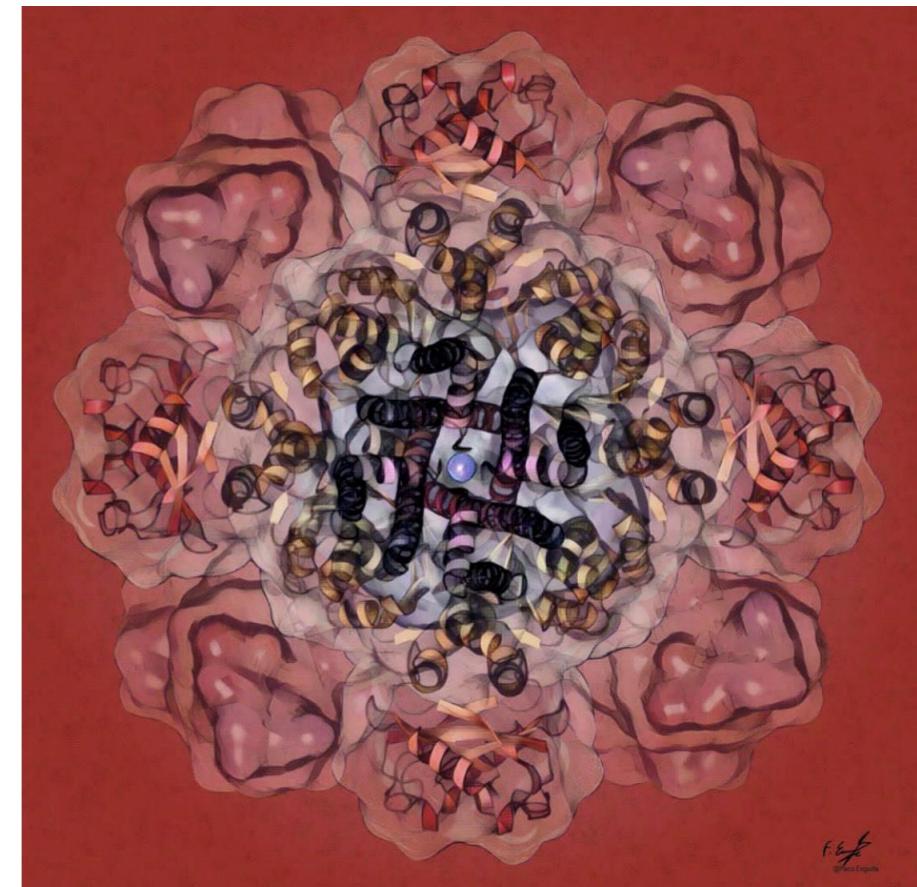
Key Findings: PersiLac3 had the highest binding affinity and greatest effect on LDPE degradation.



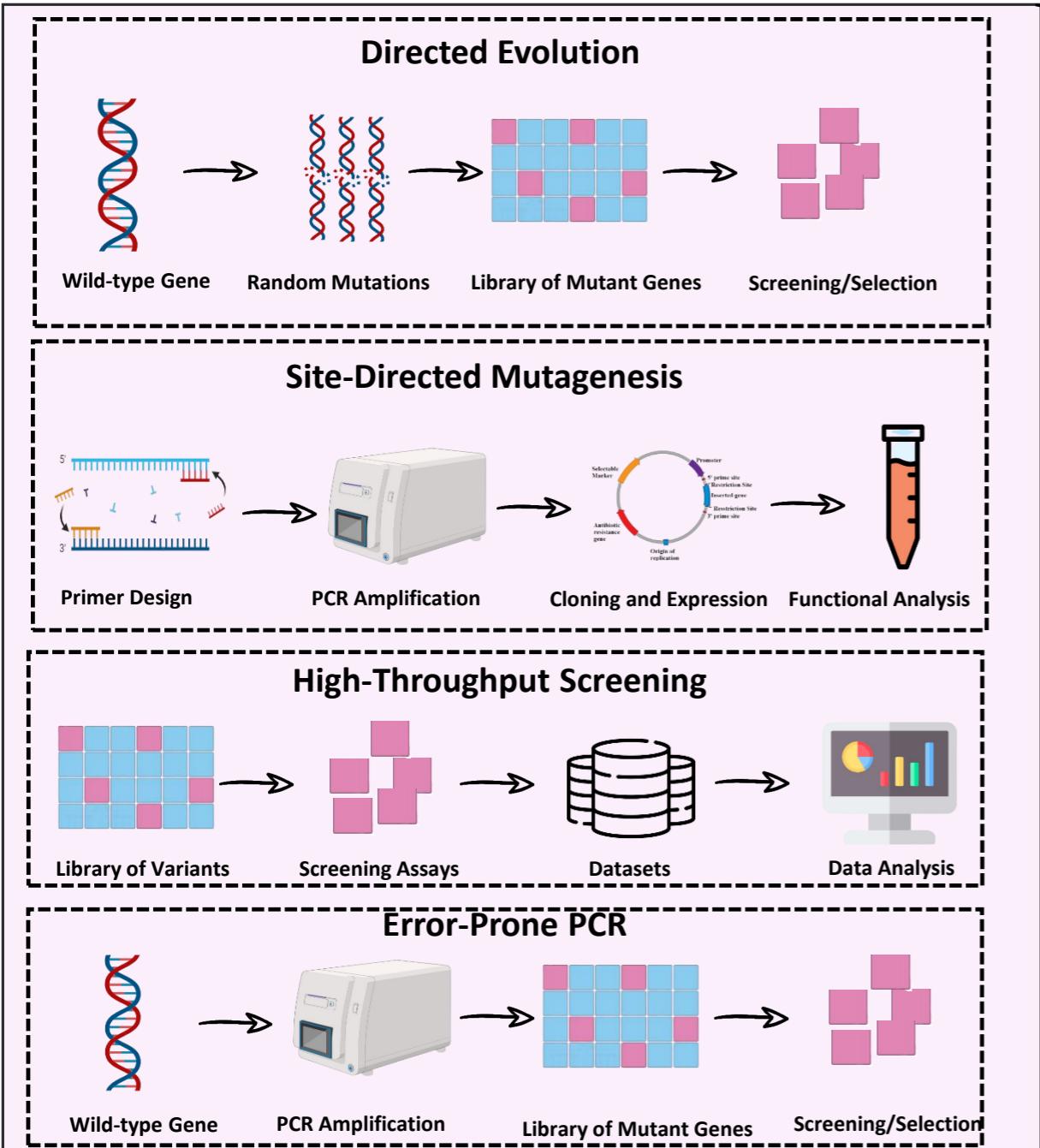
# Enzyme Engineering

Protein engineering is the process of developing new or improved proteins to enhance their performance in various industrial applications.

- ❖ 1970s: Growth with recombinant DNA technology.
- ❖ 2018 Nobel Prize for Frances Arnold's work on directed enzyme evolution.
- ❖ Improving Enzyme Characteristics for Industry
- ❖ Goal: Enhance enzyme suitability for industrial processes.
- ❖ Challenges: Predicting mutation effects.
- ❖ Benefits: Adaptation to non-aqueous solvents for biocatalysis.

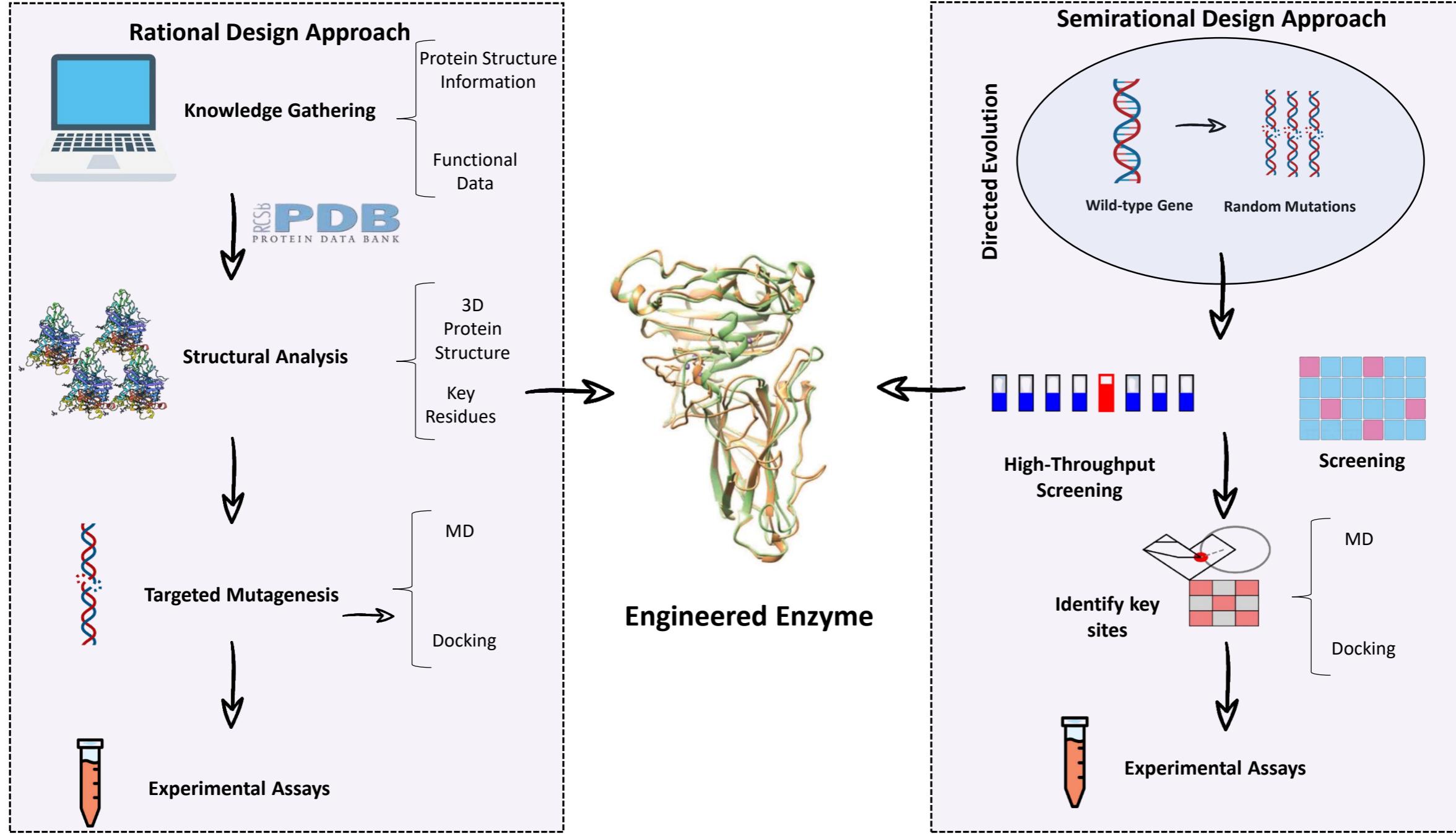


# Enzyme Engineering: Experimental Approach



Engineered Enzyme

# Enzyme Engineering: Computational Approach



# Enzyme Engineering: Computational Approach

## Structure-Guided Consensus Approach in Protein Engineering

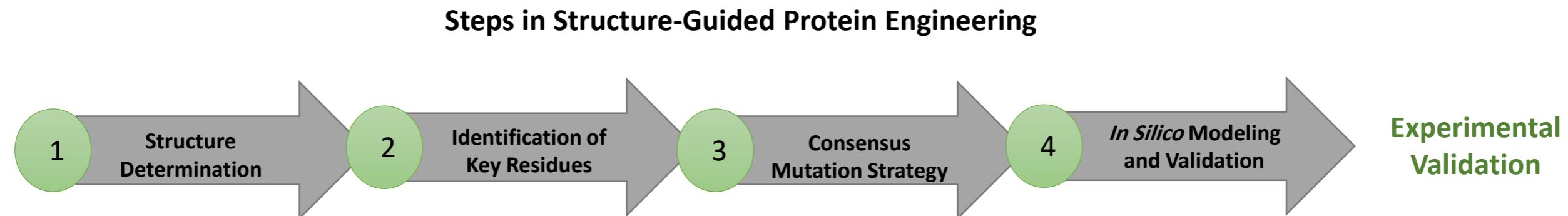
### Overview

#### Goal

To enhance protein function

#### Principle

This method integrates detailed protein structure information to identify critical regions for modification, aiming to achieve optimized protein performance.



# Enzyme Engineering: Computational Approach

## *In Silico* Screening Approach in Protein Engineering

### Overview

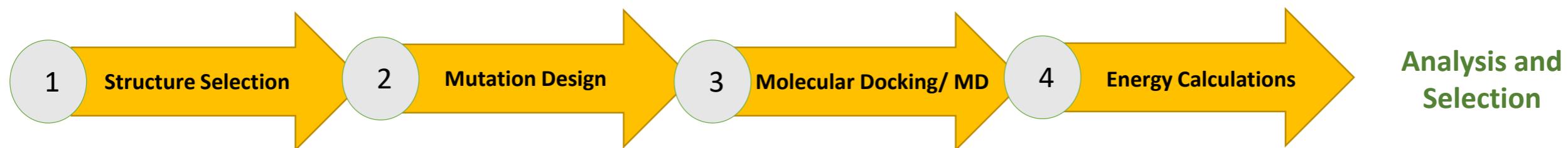
#### Goal

To predict and enhance protein functionality

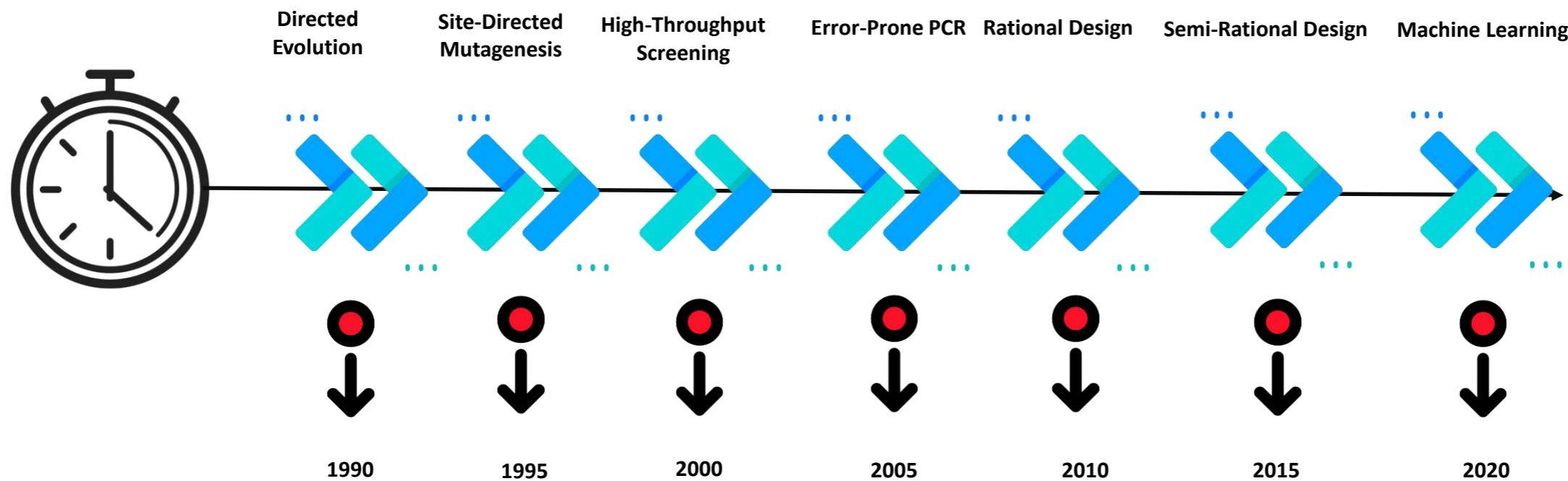
#### Principle

Utilizes computational techniques such as molecular docking, molecular dynamics simulations, and free energy calculations to screen and prioritize mutations

### Steps in Structure-Guided Protein Engineering

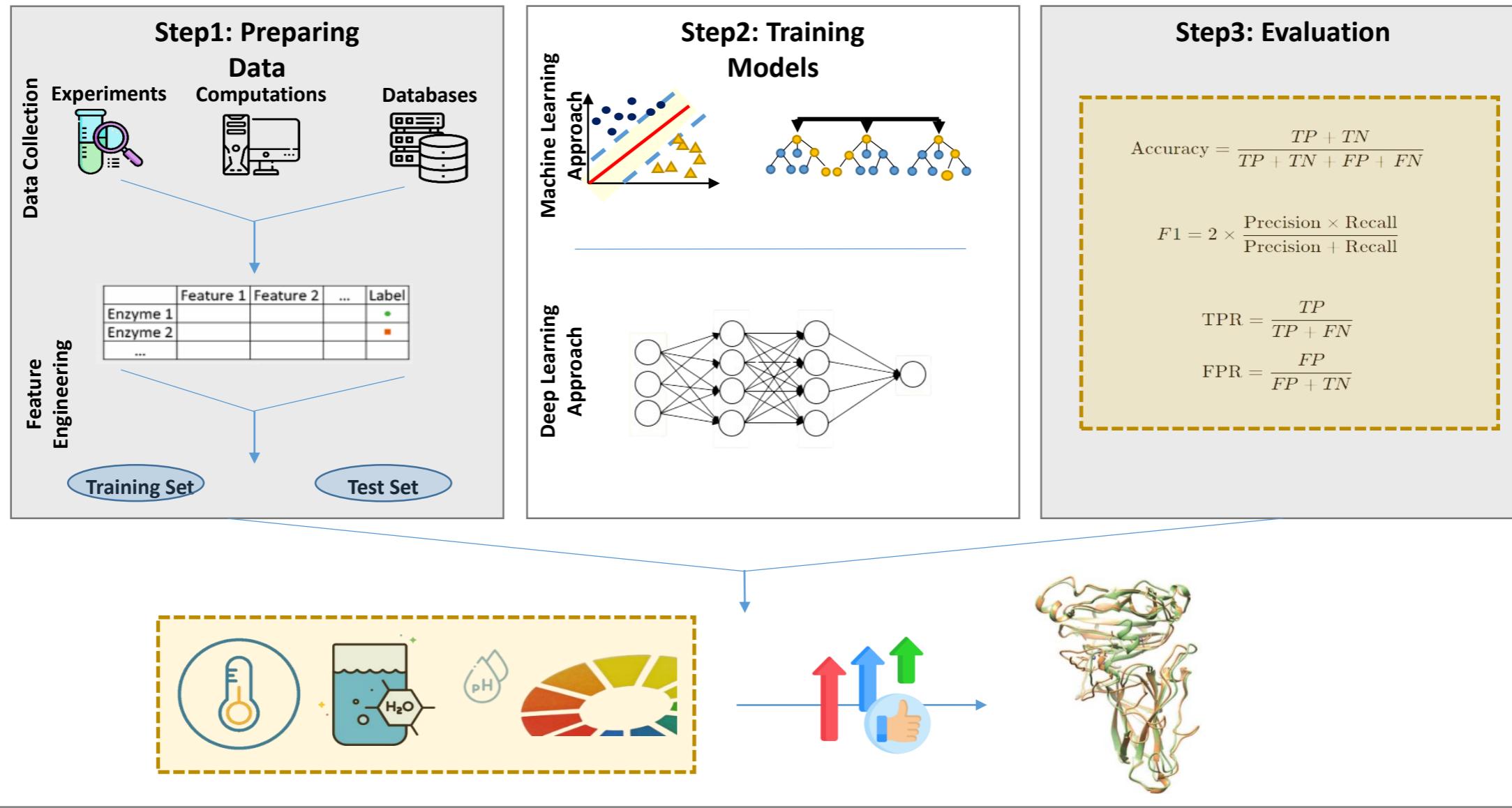


# Enzyme Engineering: Computational Approach



# Machine Learning in Protein Engineering

- ❖ ML models help in discovering, designing, and optimizing proteins by analyzing vast amounts of sequence data (145,000 PDB, 215 million protein sequences).
- ❖ Protein engineering often relies on directed evolution and fitness optimization, which ML can accelerate (10,000 to 40,000 protein families).



Research article

Open Access

## Engineering proteinase K using machine learning and synthetic genes

Jun Liao<sup>1</sup>, Manfred K Warmuth<sup>1</sup>, Sridhar Govindarajan<sup>2</sup>, Jon E Ness<sup>2</sup>,  
Rebecca P Wang<sup>2</sup>, Claes Gustafsson<sup>2</sup> and Jeremy Minshull\*<sup>2</sup>

Address: <sup>1</sup>Department of Computer Science, University of California, Santa Cruz, CA 95064 USA and <sup>2</sup>DNA 2.0, 1430 O'Brien Drive, Suite E, Menlo Park, CA 94025, USA

Demonstrate a new approach to protein engineering by combining high-throughput gene synthesis with machine learning-based design algorithms. The authors sought to enhance the activity and heat stability of Proteinase K by predicting beneficial amino acid substitutions, synthesizing these variant proteins, and then experimentally validating the most effective variants. Their method significantly reduces the number of protein variants that need to be tested, improving both the efficiency and accuracy of protein engineering.

# Case study

- ❖ Proteinase K is a serine protease enzyme widely used in molecular biology for protein digestion(28.9 kDa).

## Data Input:

- ❖ 59 Protein variants (Initial Round)
- ❖ 95 Variants (Total)
- ❖ 19-Dimensional Binary Vectors

## Key Features:

- ❖ Amino acid substitutions
- ❖ Binary representation of sequence changes

## Models Used:

- ❖ Random Forest
- ❖ SVMR (Support Vector Machine Regression)
- ❖ Lasso
- ❖ PLSR (Partial Least Squares Regression)
- ❖ MR (Matching Loss)
- ❖ LPBoostR

## Best Model:

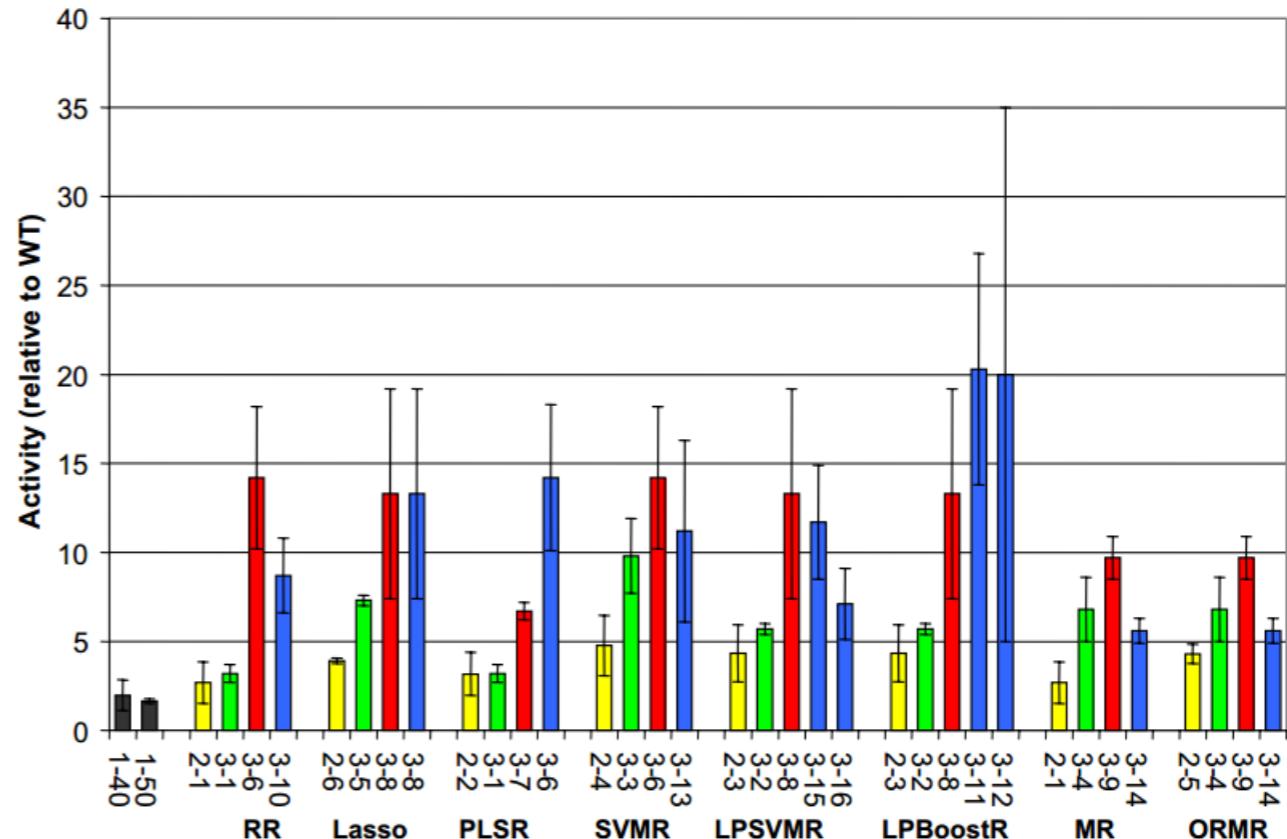
- ❖ PLSR
- ❖ LPBoostR

## Evaluation Metrics:

- ❖ R-squared ( $R^2$ )
- ❖ Mean Squared Error (MSE)

## Output:

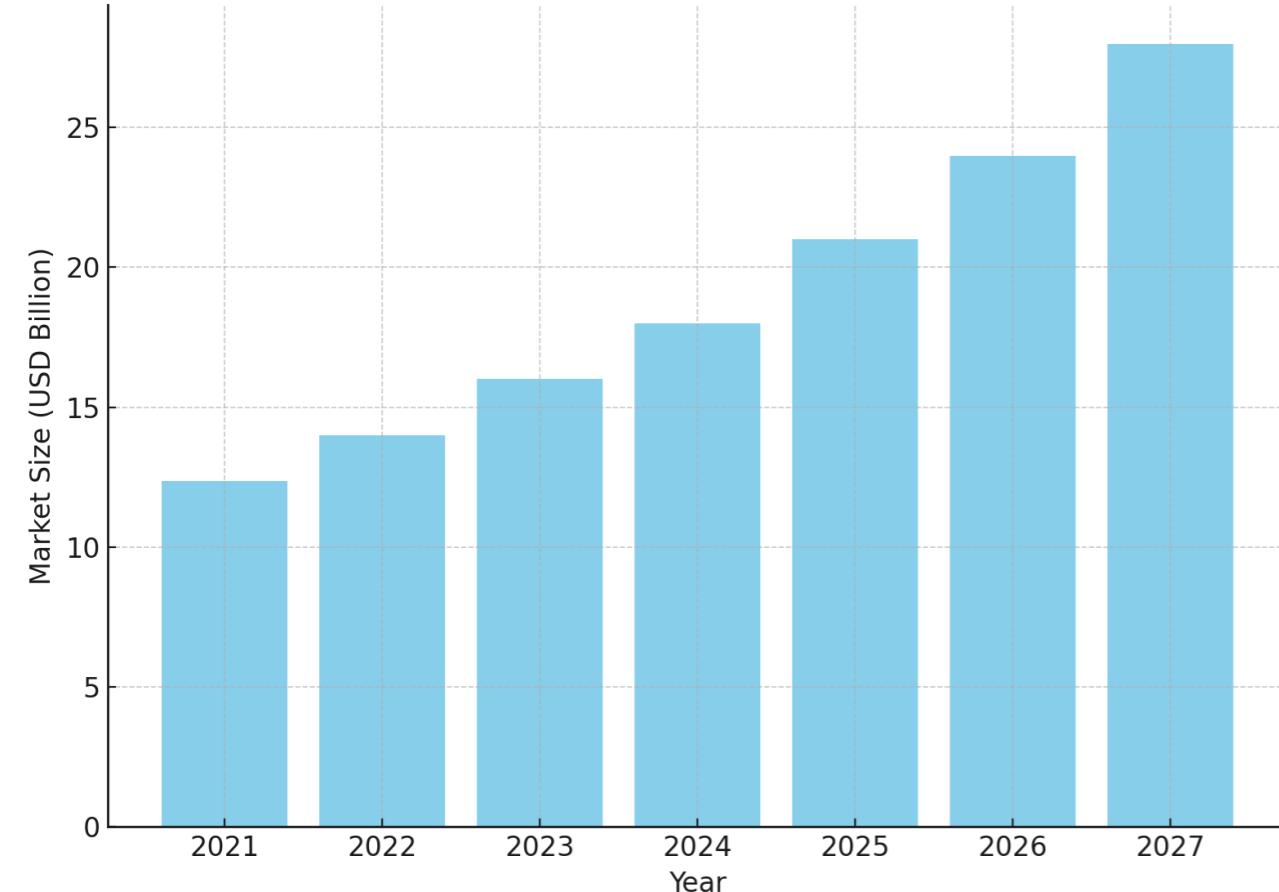
- ❖ Predicted enzyme activity
- ❖ Experimental validation (up to 20x activity improvement)



# Marketing

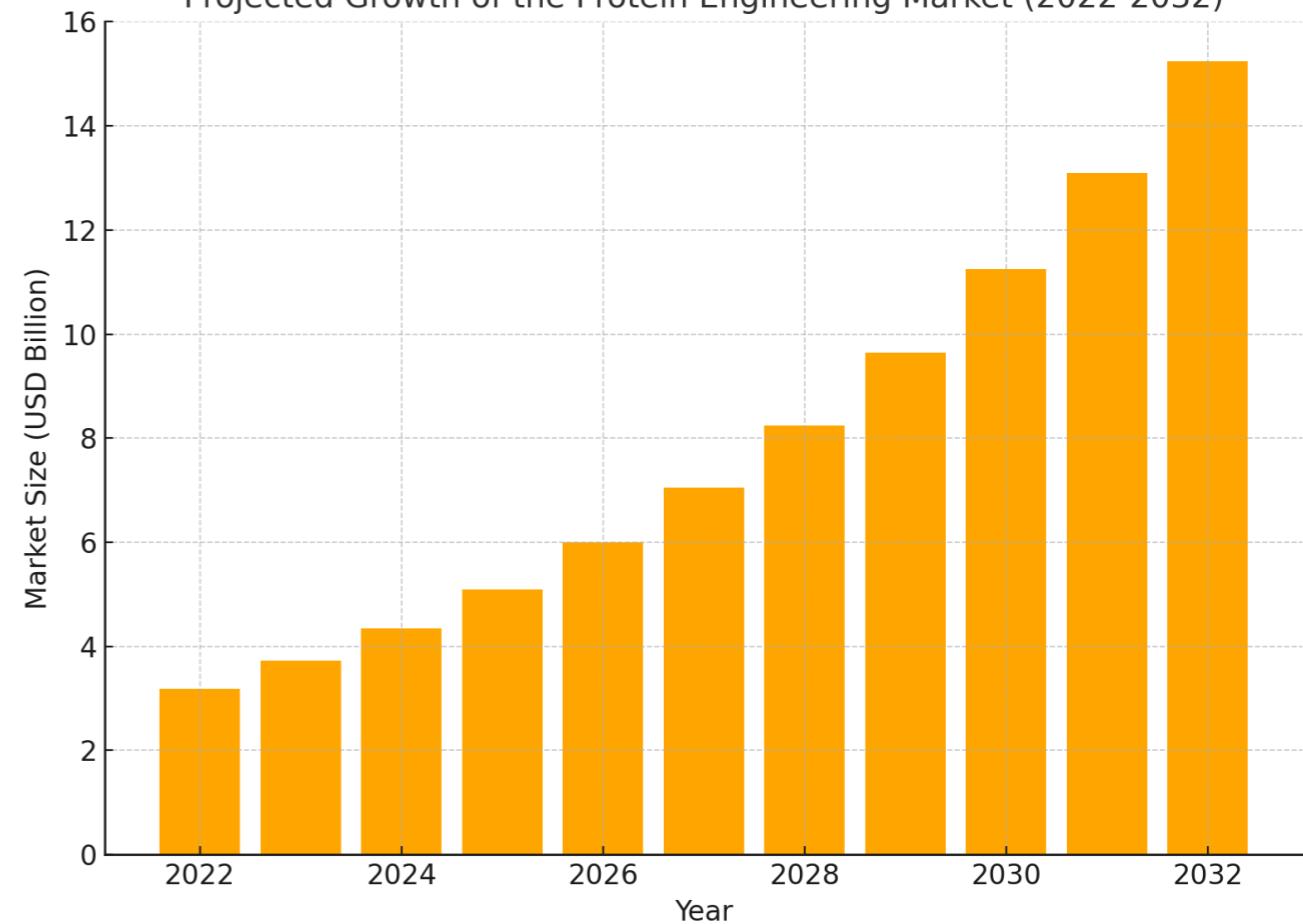
## Structural bioinformatics market

Projected Growth of the Structural Bioinformatics Market (2021-2027)



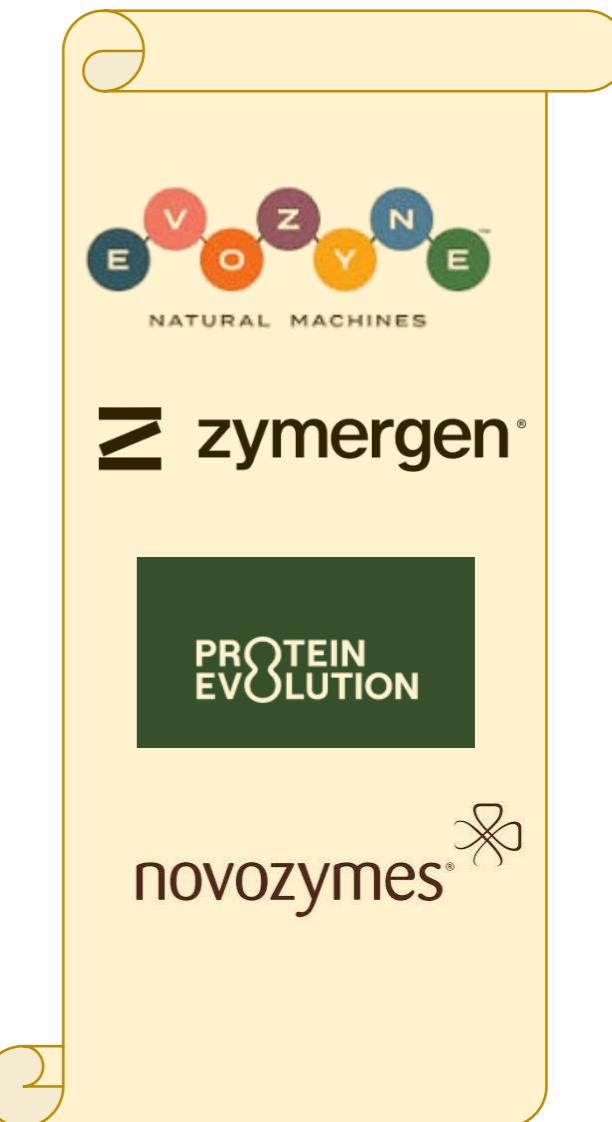
## Protein engineering market

Projected Growth of the Protein Engineering Market (2022-2032)



# Enzyme Engineering: Companies

Company Name	Description
Sana Biotechnology	Focuses on gene and cell engineering, aiming to revolutionize medicine through gene modification.
Merck	Works on protein engineering, including the development of monoclonal antibodies and other therapies.
Moderna	Involved in protein engineering, especially through mRNA vaccines and therapeutic proteins.
Generate Biomedicines	Uses machine learning to create novel proteins, antibodies, peptides, and enzymes.
AI Proteins	Specializes in designing therapeutic miniproteins using AI and synthetic biology.
BigHat Biosciences	Combines synthetic biology and machine learning for antibody discovery and engineering.
Ventus Therapeutics	Engineers proteins targeting immune system elements, like inflammasomes, for therapies.
Evozyne	Uses generative AI to design high-performance proteins with evolutionary guidance.
Ginkgo Bioworks	Involved in protein and enzyme engineering, with a focus on bio-industrial applications.
Amgen	Collaborates with Generate Biomedicines for the development of protein therapeutics.



# Current International Projects



## Problem

IgG proteases are bacterial enzymes that have been shown to cleave IgG. However, these proteases are non-specific and can trigger an immune response that affects their efficacy, stability, and redosability.

## Evozyne Solution

Design novel IgG proteases that have several points of differentiation: higher activity and stability to minimize dosage, reduced immunogenicity to enable redosability, and increased specificity to disease-causing subtypes of IgG to increase safety.

## Problem

Existing nucleases are not specific enough and can cause off-target editing (i.e., edit the genome in areas not intended), hampering its widespread use in therapeutic applications.

## Evozyne Solution

By augmenting and redesigning the nucleases created by nature, Evozyne is developing a comprehensive library of highly specific Cas nucleases that can access most of the genome. This combination of accessibility and specificity will allow previously untreatable diseases to be cured safely and effectively with our enzymes.

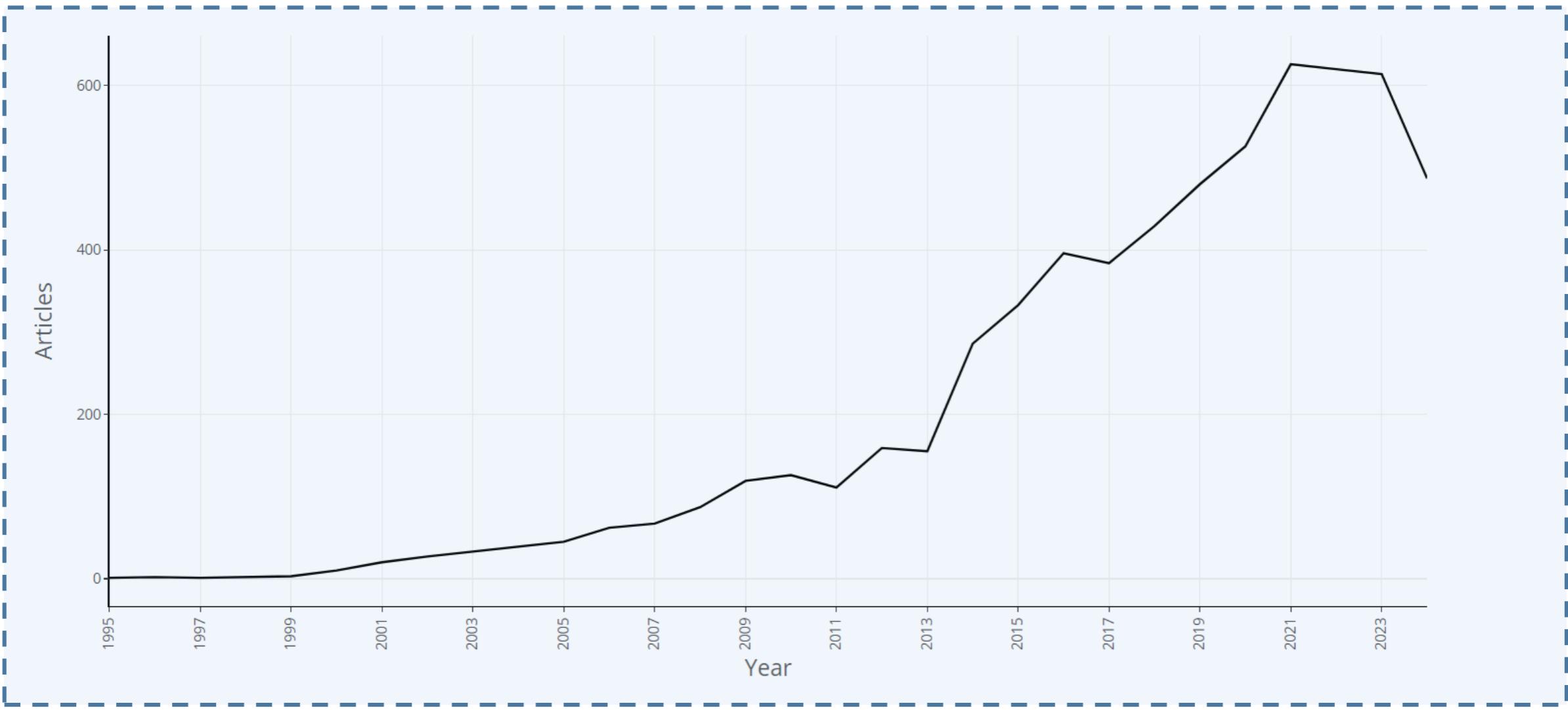
## Problem

The sugar substitute market demands natural alternatives that are healthier and more efficient than traditional sweeteners.

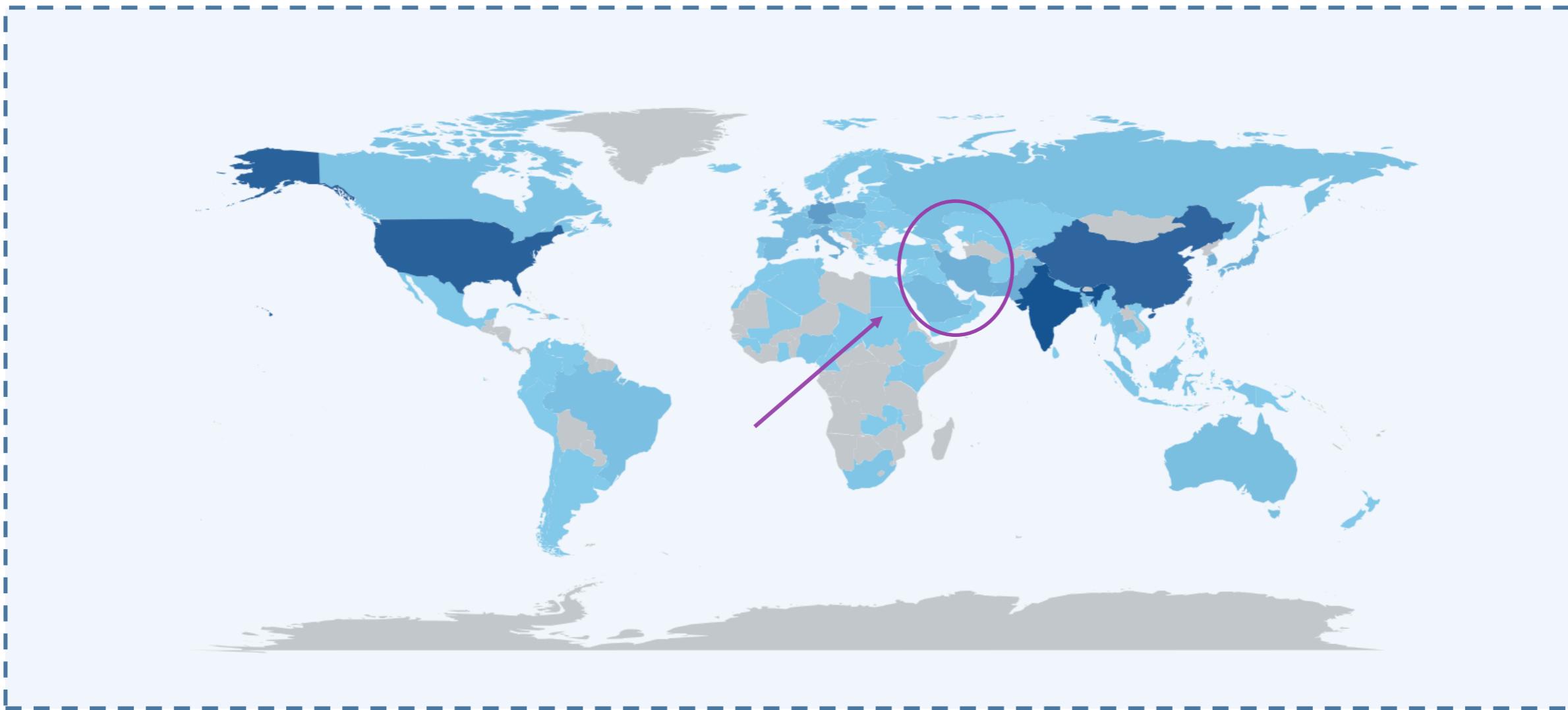
## GreenLab and Ginkgo Bioworks Solution

GreenLab is working with Ginkgo Bioworks to scale up the production of brazzein, a naturally sweet protein, by growing it inside corn kernels using proprietary technology. With Ginkgo's Plant Trait Services, Protein Services for precision fermentation, and Deployment Capabilities for efficient purification, this partnership aims to transform the sugar substitute industry. By offering a natural, highly potent alternative to sucrose, GreenLab is positioning itself as a leader in the market while minimizing production risks and costs, ensuring sustainable and scalable solutions.

# Historical Growth of Research Output by Year



# Global Distribution of Scientific Research Output



**Computational Insights into the Selecting Mechanism of  $\alpha$ -Amylase Immobilized on Cellulose Nanocrystals: Unveiling the Potential of  $\alpha$ -Amylases Immobilized for Efficient Poultry Feed Hydrolysis, Bioconjugate Chemistry, 2023**

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

**Engineering proteinase K using machine learning and synthetic genes, BMC Biotechnology 2007**

## **References**

