# Abstract

We used 45730 observations from the 2002 Critical Assessment of Structure Prediction 5 Physiochemical Properties data set to train an optimized random forest regression model and deep neural network regression models using properties of amino acid residues to predict root-mean-square deviation of atomic positions (RMSD) for structural models. We investigate which physiochemical characteristics of amino acid residues are linked to higher error in the predicted tertiary structure and discuss the order of prediction dependence inherent to RMSD of structural models. Given the lack of order in the dataset, we found gradient boosted tree methods to most accurately predicted structural error which heavily weight polar surface area of residues.

Prediction of RMSD of Atomic Position in CASP-5 Protein STructure Models

Comparison of Gradient Boosted Regression Tree, Feedforward , and GMDH Deep Neural Network Methods

EE

Evan Domingos - 2023

domingos\_evan@wheatoncollege.edu

Table of Contents

Abstract [1](#_Abstract)

Introduction [4](#_Introduction)

Problem Statement

Basics of Protein Structure & Protein Folding [5](#_Basics_of_Protein)

Calculation of RMSD [5](#_Structural_Alignment_and)

Pairwise Structural Alignment [5](#_Structural_Alignment_and)

CASP 5.9 Dataset [5](#_CASP_5.9_Dataset)

Preliminary Discussion [4](#_Preliminary_Discussion)

Methods [4](#_Methods)

Linear Regression Models [5](#_Basics_of_Protein)

Random Forest Regression Tree [5](#_Basics_of_Protein)

Feedforward Neural Networks [5](#_Basics_of_Protein)

GMDH Neural Networks [5](#_Basics_of_Protein)

Gradient Boosted Regression Tree [5](#_Basics_of_Protein)

Results [4](#_Results)

Discussion [4](#_Results)

Conclusion [4](#_Conclusion)

Appendix [4](#_Appendix)

Bibliography [4](#_Appendix)

# Introduction

## Problem Statement

Proteins are vital to the functions of human cells, and they are responsible for performing a variety of tasks. There are tens of thousands of unique proteins within a typical human cell. Proteins fold into functional shapes and their different structures allow them to perform diverse roles. However, predicting the three-dimensional structure of a protein from its amino acid sequence remains a challenging task. Computational methods like machine learning algorithms have proven extremely well fitted for the task of protein folding predictions but it is crucial to identify and discard predicted models with structural errors, as relying on incorrect models can lead to erroneous conclusions about protein function and hinder the development of effective therapeutics. Some predictive models include predicted structural error factors to adjust predictions, similar to gradient boosting. We aim to construct a model to predict the structural errors in protein folding models from the 2002 CASP-5 competition.

## Basics of Protein Structure & Protein Folding

Proteins are large biomolecules vital for performing functions within organisms like the replication of DNA, metabolizing enzymes, and structuring of cells. Proteins are formed from long chains of organic compounds called amino acids. Amino acids are formed around a single carbon atom called an alpha-carbon. Carbon has a valence of four so it may form four single bonds. In an amino acid, the alpha carbon is bonded to an amino group (NH2), a carboxyl group (COOH), a single hydrogen atom, and a variable element called the “R” group.

Two amino acids can form a peptide bond between the carboxyl group of one amino acid and the amino group of the other. A single water molecule (H2O) is formed from the expelled (H-) and (H0+), leaving a hydrophobic bond between the two amino acids called a peptide bond. The combined amino acids are called a polypeptide.

A picture containing chart

Description automatically generated

Any number of amino acids can bond to that structure, aligning the carboxyl group to the amino group of each successive amino acid resulting in a chain referred to as the primary structure of a protein. Two or more combined amino acids forming a peptide are also called a residue. There are 20 organically common amino acids that can be categorized by the chemical properties of their R-groups; polarity and electrical charge. The physical interaction of the R-group and the hydrogen bonds linking the amino acids influences the angle of the bond, contorting the flat primary structure into either a helical structure or a pleated sheet, named alpha helices and beta sheets respectively. This is referred to as the secondary structure. The tertiary structure is determined by the interaction of R-groups within the polypeptide chain.

These interactions include hydrogen bonding, ionic bonding, and hydrophobic interactions. The R-groups of amino acids that are hydrophilic (polar) tend to be located on the protein's surface, while those that are hydrophobic (non-polar) are generally buried in the structure's interior, away from the water. These physical interactions between the R groups of amino acids determine the protein's final three-dimensional shape, the tertiary structure which is ultimately responsible for determining the function of the resulting protein.

Primary Structure: Sequence of bonded amino acids (residues) forming a polypeptide

Secondary Structure: 3D structure from the interaction of R-groups and hydrogen bonds

Tertiary Structure: Contortions from the interaction between R-groups

Given the importance of protein function and structure in biochemistry for the development of drugs among other applications, the prediction of structure from amino acid sequences is of immense interest. Since the structures of proteins are determined by their amino sequences, protein structure prediction or protein folding has shown to be a powerful use-case for machine learning techniques and neural networks in particular. The use of machine learning algorithms in protein prediction has significantly reduced the time and cost required for the experimental determination of protein structures, making it a valuable tool in biochemistry research. Just 200,000 protein structures have been verified in the 130 years since x-ray crystallography was invented (Source: Protein Data Bank, 2023). In a single year, a neural network developed by Google’s DeepMind was able to produce and publish predictions for over 200 million known amino sequences with impressive results on verified structures.

This model named Alpha was developed for the Critical Assessment of Structure Prediction (CASP) which is a biennial worldwide experiment to evaluate the efficacy of computational protein structure prediction methods developed by academia and industry teams. CASP allows researchers from around the world to test their prediction methods on a set of protein sequences whose structures have not been determined experimentally or whose verified structures have not been released.

The scoring data from the 2002 CASP 5 competition was published to the UCI ML Database which was used for this project. The data contains residue prediction scores for models submitted for the competition, measuring the deviation in position between the predicted structure and the aligned verified protein structure.

## Calculation of RMSD

Protein prediction models are evaluated on the residuals of the predicted structure and the verified structure. Results in the CASP 5.9 dataset are evaluated using root mean squared deviation of atomic position (RMSD).

Both modeled and verified structures have the same spatial geometry. The structure is indexed by the amino sequence. For the purpose of prediction scoring, the sequences are identical.

Example:



Source: *Protein Data Bank*

The alpha carbon for a given residue is treated as the location for that amino acid and each residue in the structural model has XYZ coordinates. Generally speaking, the XYZ coordinates treat the first amino residue of the sequence as the 0 coordinate. Thus, the location of following amino acids are dependent on the location of the preceding residue.

Root mean squared deviation is calculated at each residue for the difference between the predicted and verified locations of the alpha carbon.

The general equation for RMSD of atomic position in an XYZ coordinate system is:

where n is the number of points in the measurement (residues in the structure where v is the verified location and w is the predicted location. In the context of this paper, observations are scores for single residues thus we can consider that n = 1 for all measurements and the formula for RMSD is equivalent to the distance formula in a 3-dimensional cartesian coordinate system.

Again, v is the verified alpha carbon location and w is the predicted location. RMSD is reported in Angstroms (A) where 1A = 10E-10 meters.

## Pairwise Structural Alignment

**Structural geometry in a 2-dimensional coordinate system**



**Pairwise Structural Alignment**





## CASP 5.9 Dataset

The CASP 5.9 Dataset contains 45730 observations of scored results from the 2002 CASP 5 competition. 67 newly validated proteins where used to score structure predictions of the submitted models. This dataset is ***not*** the structural predictions themselves. Each observation represents a prediction for a single amino acid and contains physiochemical properties from the prediction as well as RMSD, the locational residual between the

It is not clear the number of unique residues contained within the score (multiple models making predictions for the same protein) but the CASP 5 validation set contains a total of 14882 residues, suggesting some residues must be repeated.

The dataset has 9 features and the label RMSD. The 9 features are all continuous numerical variables defined as the following:

F1 - Total surface area (Angstroms

F2 - Non-polar exposed area

F3 - Fractional area of exposed non-polar residue.

F4 - Fractional area of exposed non-polar part of residue

F5 - Molecular mass weighted exposed area.

F6 - Average deviation from the standard exposed area of residue.

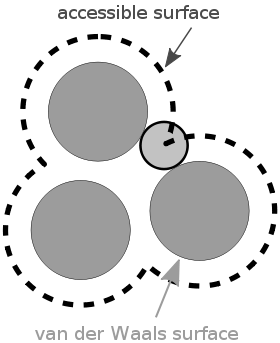
F7 - Euclidean distance.

F8 - Secondary structure penalty.

F9 - Spatial Distribution constraints (N,K Value).

Exposed or accessible surface area (ASA) is the surface area of the protein that is accessible to a solvent, water in this application. One method of calculating ASA is the Shrake-Rupley algorithm which draws a mesh of points around each atom of the residue with a radius equal to the Van der Waals radius (conceptually the surface of the atom) plus the radius of the solvent (water is 1.4Å.) [Shrake Rupley Algorithm]

Visualization of the



# Preliminary Discussion

From the Prediction Center website (CASP5)

Example:

Model name T0201TS012\_2u has the following components:

T0201 target number

TS Tertiary Structure (3D atoms coordinates) prediction

012 prediction group 12

2u model index 2, UNREFINED set of coordinates

# Methods

## Regression Tree

## Linear Regression

## Feedforward Neural Network

## GMDH Neural Network

## Gradient Boosted Regression Tree

Chart

Description automatically generated

# Results

# Discussion

# Conclusion

# Appendix