



Edwin Germán Maldonado Távara

**Investigation of the relative importance of
features used in protein structure prediction: a
machine learning approach.**

Tese de Doutorado

Thesis presented to the Programa de Pós-graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica.

Advisor : Prof. Marley Maria B. R. Vellasco
Co-advisor: Dr. Bruno A. C. Horta
Co-advisor: Dr. Fábio Lima Custódio

Rio de Janeiro
August 2017



Edwin Germán Maldonado Távara

**Investigation of the relative importance of
features used in protein structure prediction: a
machine learning approach.**

Thesis presented to the Programa de Pós-graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica. Approved by the undersigned Examination Committee.

Prof. Marley Maria B. R. Vellasco

Advisor

Departamento de Engenharia Elétrica – PUC-Rio

Dr. Bruno A. C. Horta

Co-advisor

Universidade Federal de Rio de Janeiro – LABBMOL/UFRJ

Dr. Fábio Lima Custódio

Co-advisor

Laboratório Nacional de Computação Científica – LNCC

Prof. André Vargas Abs.

Universidade do Estado do Rio de Janeiro – UERJ

Profa. Karla Figueiredo

Universidade do Estado do Rio de Janeiro – UERJ

Prof. Laurent Emmanuel Dardenne

Laboratório Nacional de Computação Científica – LNCC

Dr. Marcos Henrique de Pinho Maurício

Departamento de Engenharia de Materiais – PUC-Rio

Prof. José Eugenio Leal

Vice Dean of Graduate Studies

Centro Técnico Científico – PUC-Rio

Rio de Janeiro, August the 30th, 2017

All rights reserved.

Edwin Germán Maldonado Távara

Graduated in Systems Engineering at UNT (National University of Trujillo - La Libertad, Perú) in 2002. Did his master degree at PUC-Rio, specializing in the application of optimization and machine learning methods in Bioinformatics. Dedicated to full-time research at PUC-Rio, Brazil.

Bibliographic data

Maldonado Távara, Edwin Germán

Investigation of the relative importance of features used in protein structure prediction: a machine learning approach. / Edwin Germán Maldonado Távara; advisor: Marley Maria B. R. Vellasco; co-advisores: Bruno A. C. Horta, Fábio Lima Custódio. – Rio de Janeiro: PUC-Rio, Departamento de Engenharia Elétrica, 2017.

v., 25 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Inteligência Computacional Aplicada – Teses. 3. Predição de Estrutura de Proteínas;. 4. Avaliação da Qualidade de Proteínas;. 5. Seleção de descritores;. 6. Importância relativa de descritores;. 7. Aprendizado de Máquina;. 8. Redes Neurais;. 9. Algoritmos Genéticos.. I. Vellasco, Marley. II. Horta, Bruno. III. Custodio, Fabio. IV. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. V. Título.

CDD: 620.11

To my parents, for their support
and encouragement.

Acknowledgments

I would like to first thank my advisor ...

Then I wish to thank ...

Abstract

Maldonado Távara, Edwin Germán; Vellasco, Marley (Advisor); Horta, Bruno (Co-Advisor); Custodio, Fabio (Co-Advisor). **Investigation of the relative importance of features used in protein structure prediction: a machine learning approach..** Rio de Janeiro, 2017. 25p. Tese de doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Proteins are important because they determine different functions in living cells. These functions, are directly related to its three-dimensional structure. To determine the protein three-dimensional structure have been developed several experimental methods. However, these methods are costly and time-consuming. Therefore, computational methods have been developed for protein tertiary structure prediction. These methods generate a set of several candidate's models known as decoys. In this set, identify the model or subset of models that are more close to the protein native structure is a challenging task. To perform this task two types of methods have been proposed. The first types of methods use the native structure to evaluate the decoys quality (RMSD,GDT-TS,TM-Score,MaxSub). In absence of the native structure, methods based scoring functions have been proposed (physics-based potential functions, statistical potential functions, consensus-based functions, machine learning algorithms). Machine learning methods, evaluate the models quality using a subset of features. The advantage of these methods is their power to identify hidden relationships between the selected features, this task is difficult to achieve with the other methods. On the other hand, the main disadvantage is that these methods perform a non-automatic feature selection, this could be a factor that limits the quality assessment of the decoys models. Given these drawbacks, this work proposes a model for the automated feature selection, this selection is performed through different types of features. Additionally, this work introduces a new wrapper method to calculates the relative feature importance called (SWtoFIC). Finally, the model provides the GDT-TS score prediction for assessing the quality of the protein decoys.

Keywords

Protein Structure Prediction; Protein Quality Assessment; Feature Selection; Features Relative Importance; Machine Learning; Neural Networks; Genetic Algorithm.

Resumo

Maldonado Távora, Edwin Germán; Vellasco, Marley; Horta, Bruno; Custodio, Fabio. **Investigação da importância relativa de features usados na predição da estrutura de proteínas: Uma abordagem usando Aprendizado de Máquina.** Rio de Janeiro, 2017. 25p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

As proteínas são importantes pois estas determinam as funções das células vivas. Estas funções estão diretamente relacionadas com sua estrutura terciária. Para determinar a estrutura terciária de proteínas tem sido desenvolvidos muitos métodos experimentais. Sin embargo, estes métodos são custosos e demorados. Por isso, métodos computacionais foram desenvolvidos para a predição da estrutura terciária da proteína. Esses métodos geram um conjunto de vários modelos candidatos conhecidos como decoys. Neste conjunto, identificar o modelo ou o subconjunto de modelos mais próximos da estrutura nativa da proteína, é uma tarefa desafiadora. Para realizar esta tarefa, foram propostos dois tipos de métodos. O primeiro grupo de métodos utilizam a estrutura nativa para avaliar a qualidade dos decoys (RMSD, GDT-TS, TM-Score, MaxSub). Na ausência da estrutura nativa, foram propostas métodos baseados em funções de pontuação (funções de potencial, funções de potencial estatístico, funções baseadas em consenso, algoritmos de aprendizado de máquina). Os métodos de aprendizado de máquina, avaliam a qualidade dos modelos decoys usando um subconjunto de descritores. A vantagem desses métodos é a característica de identificar relações ocultas entre os descritores selecionados, esta tarefa é difícil de alcançar com os outros tipos de métodos. Por outro lado, estes métodos tem a desvantagem de realizar uma seleção não automática de descritores, isto pode ser um fator limitante na avaliação da qualidade dos decoys. Dadas estas desvantagens, este trabalho propõe um modelo para a seleção automática de descritores, esta seleção é realizada usando diferentes tipos de descritores. Além disso, este trabalho introduz um novo método para o cálculo da importância relativa dos features chamados SWToFIC. Finalmente, o modelo fornece a predição do GDT-TS score para a avaliação da qualidade dos decoys de proteína.

Palavras-chave

Predição de Estrutura de Proteínas; Avaliação da Qualidade de Proteínas; Seleção de descritores; Importância relativa de descritores; Aprendizado de Máquina; Redes Neurais; Algoritmos Genéticos.

Table of contents

1	Introduction	13
1.1	Motivation	13
1.2	Objectives	16
1.3	Contributions	16
1.4	Work Description	17
1.5	Work Organization	17
2	Review	18
2.1	Hematite	18
2.1.1	Martite	18
2.1.1.1	Globular	18
3	Conclusions	20
	Bibliography	21
A	Published paper	25

List of figures

List of tables

Table 2.1	Main morphologies of hematite.(14)	19
-----------	------------------------------------	----

Like all other arts, the Science of Deduction and Analysis is one which can only be acquired by long and patient study, nor is life enough to allow any mortal to attain the highest possible perfection in it. Before turning to those moral and mental aspects of the matter which present the greatest difficulties, let the inquirer begin by mastering more elementary problems.

Sir Arthur Conan Doyle, *Sherlock Holmes, A Study in Scarlet*.

List of Abbreviations

ADI – Análise Digital de Imagens

BIF – *Banded Iron Formation*

... – ...

1

Introduction

1.1

Motivation

In nature, there are 20 different natural amino acids, each of them exhibiting different physicochemical properties (e.g. size, charge, hydrophobicity)(1)(2). Depending on their polarity, amino acids vary in their hydrophilic or hydrophobic character, which is crucial for the formation of more complex structures. A peptide is a molecule composed of two or more amino acid residues linked through a covalent chemical bond called the peptide bond. This peptide bond is formed when the carboxyl group of a residue reacts with the amino group of another residue, releasing one water molecule. Large peptides are generally called polypeptides or proteins(3)(4). Proteins perform a variety of functions in living organisms such as catalysts of specific reactions, structural components of cell membranes, antibodies, carry oxygen in the blood and are part of chromosome material. On the other hand, proteins control the regulation and reproduction of living things and are thus essential for life.

The structure of proteins can be divided into four levels(1)(2) (a) primary structure, (b) secondary structure, (c) tertiary structure and (d) quaternary structure. The primary structure corresponds to the sequence (linear order) of the amino acids forming a given protein(1)(2)(4)(5). The beginning of the primary structure corresponds to its N-terminal and the end of its primary structure is the C-terminal region. The secondary structure is defined by characteristic structural arrangements that are the consequence of hydrogen-bond patterns involving the backbone atoms (the amide chemical function that forms the main chain of the peptide)(2). The most common regular secondary structures are α -helices and β -sheets(6), which are usually highly stable and constitute key elements in the three-dimensional (3D) structure of proteins. The tertiary structure of a protein is represented by the distribution of secondary structures in a 3D space.

The three-dimensional shape assumed by a protein is also called its native, functional or folded structure. The native structure is the lowest free-energy conformation under physiological condition, which can be viewed as a consequence of the compromise between enthalpy (Intra- and intermolecular interactions) and entropic (related to the number of accessible states) contributions(1)(2)(4)(7). The knowledge of the tertiary structure of a protein is usually of great importance in biology as it permits: the analysis and prediction of protein function in cells; the identification of active sites and binding sites of a receptor; or the identification of a site of recombination to the action of another protein(2). Finally, the quaternary structure is defined for proteins having multiple domains of tertiary structure and is defined by the spatial arrangement of these domains.

One of the main challenges in structural bioinformatics concerns the determination of protein tertiary structures. Determining the tertiary structure of proteins are experimentally and computationally expensive and time consuming(8). The difficulty in determining 3D structures of proteins has generated a large imbalance between the number of known amino acid sequences and the number of 3D structures of proteins. In spite of a large number of known protein sequences, only a small fraction possess associated 3D structures. This emphasizes the need of computational methods for the prediction of protein tertiary structures.

A number of computational methods have been proposed as a solution to the problem of protein structure prediction (PSP)(9)(10)(11)(12). These methods can be divided into four classes (13): (a) fold recognition and threading methods(14)(15)(16)(17); (b) comparative modeling methods and sequence alignment strategies (18)(19) (c) first principle methods with database information(20)(21); (d) first principle methods without database information(10). The first three groups of methods are capable of predicting the tertiary structure of proteins quickly and efficiently when template structures or folding libraries are known(22). However, the last group (first principle methods without database information), which is referred to as *ab initio*, is based on first principles and do not rely on database information and produces new folds through computer simulation of physicochemical properties of the process of protein folding in nature. All these techniques generate a large number of candidate models known as decoys (23). Determine the models quality assessment is a challenging problem in structural biology and is an important area in bioinformatics(24). Several model quality assessment methods(MQA) have been developed since CASP7 (7th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction)(24)(25).

Over the past years, several techniques have been developed, these techniques can be classified into two big groups: When the native protein structure is known and when is not. The first group of techniques evaluates the quality of a decoy through different measures of similarity such as Root-Mean-Square Deviation (RMSD) (26), Global Distance Test Total Score (GDT_TS) (27), Template-Modeling Score (TM-Score) (28) and Maximal Substructure (Max-Sub) (29). The second group of methods uses scoring functions that allows one to discriminate between high and low-quality models. These scoring functions can be classified as follows: (a)physics based potential functions; (b)statistical potential functions; (c)consensus-based functions; and (d)machine learning algorithms. Physics based functions calculate the potential energy associated with a given protein model and may incorporate also the interactions with the surrounding solvent(30)(31). Such methods usually consume a significant amount of computational time and are not very sensitive to small atomic changes. Statistical potential functions evaluate the quality of the models based on statistical analysis of structural attributes extracted from a known protein structure database(32)(33). However, these methods only consider average properties of known protein structures and therefore have limited power to discriminate and rank structural models. Consensus-based functions have a good performance when many models in the database are similar to the native structure. However, if the database is composed of only low-quality models, these methods tend to show less performance than knowledge-based approaches(34)(35). Machine learning algorithms, such as support vector machine (SVM), neural networks (NN) and random forest (RF), evaluate the quality of the models in relation to certain features or attributes(36)(37). These features are extracted from the sequence and structure of decoy models and are used as inputs to evaluate their quality. The great advantage of these methods is that they can consider a large number of features. These features may be considered simultaneously, and the hidden relationships between them can be taken into account, which is very difficult to achieve with the other methods described.

Features can be extracted from the secondary structure (for example, average of amino acids forming alpha helix structures). Features based on Euclidean distance are also widely used (for example, distance between the amino acids and solvent distance between aliphatic amino acids). Features can also be calculated based on the physicochemical properties of protein models, such as the number of aliphatic hydrophobic amino acids and the solvent accessibility on different levels(38)(39)(40). Furthermore, it is also possible to represent the protein as a network structure(41). Using such a representation it

is possible to calculate features such as: (a) number of noncovalent interactions (defined by the number of sides on the network); (b) set of nodes connected with maximum number of amino acids; (c) average cluster coefficient of the network; and (d) average cluster coefficient of the largest cluster.

Given this wide range of possible features, many studies select an initial set to be used as inputs for the machine learning model. This selection is performed in an arbitrary fashion, which implies in biased results and, therefore, limits the prediction power of the method.

Therefore, it is highly desirable to obtain an unbiased set of relevant features in an automatic way. This could be achieved by a computational method capable of, in a first step, perform automatic feature selection identifying the most important ones from a large pool of available features and, in a second step, evaluate the relative importance of each feature within this reduced subset.

1.2 Objectives

The main objective of this study is to calculate the importance of different types of features in protein quality assessment. The specific objectives are:

- (i) To propose a model that is able to evaluate the relative importance for an optimal subset of features;
- (ii) To use this model to predict quality of decoy models towards their native structures.

1.3 Contributions

The main contribution of this work is the proposal of a strategy that uses a hybrid model, based on a genetic algorithm and a neural network, for the calculation of the characteristics (features). This model allows specifically:

- (i) Calculate three different scores for each of the features used in this proposal;
- (ii) Calculate the similarity of decoys models relative to its native structure.

1.4

Work Description

The development of this study was done in the following steps:

- (i) Study of the main technical similarity in the literature such as RMSD, TM-score, GDT-TS, MaxSub.
- (ii) Study of machine learning techniques to solve this problem (Neural Networks, SVMR, Random Forest etc.).
- (iii) Search of features commonly used in machine learning models described above
- (iv) Pre-processing decoys database, which is necessary for the feature extraction (calculation).
- (v) Hybrid model definition (GA + NN), for the calculation of features importance.
- (vi) Implementation (Matlab and Python) and testing of the proposed model in programming languages, to adjust the proposed model.

1.5

Work Organization

The proposal is divided into four additional chapters. Chapter 2 presents a summary of the theoretical foundations necessary for understanding this work, which will consider issues such as amino acids, proteins and its different levels of organization, methods for predicting the 3D protein structure, protein structure similarity metrics, and machine learning methods to predict the quality of decoys. Finally, the chapter shows the different features used to predict the protein quality assessment. In Chapter 3, the model for the protein quality assessment calculation will be presented in details, emphasizing the calculation mechanisms of the importance of the features. The chapter 4 describes and analyzes the preliminary results obtained until now. Finally, Chapter 5 presents the discussions and conclusions on the proposed model and the results obtained.

2

Review

This is the second chapter...

In this chapter, let's have a nice table:

2.1

Hematite

A hematita é o mineral de ferro mais importante devido a sua alta ocorrência em vários tipos de rochas e suas origens diversas.(30) A composição química deste mineral é Fe_2O_3 , com uma fração mássica em ferro de 69,9% e uma fração mássica em oxigênio de 30,1%.(31)

...

2.1.1

Martite

A hematita é o mineral de ferro mais importante devido a sua alta ocorrência em vários tipos de rochas e suas origens diversas.(30) A composição química deste mineral é Fe_2O_3 , com uma fração mássica em ferro de 69,9% e uma fração mássica em oxigênio de 30,1%.(31)

...

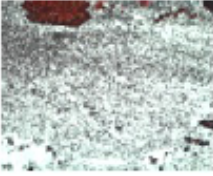

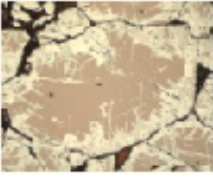
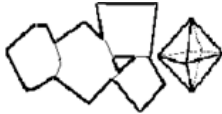
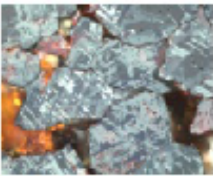

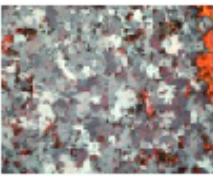

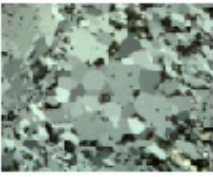
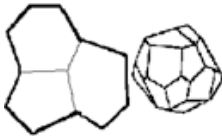
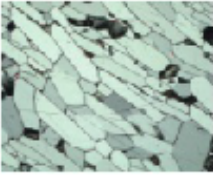

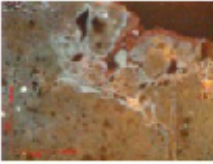

2.1.1.1

Globular

A hematita é o mineral de ferro mais importante devido a sua alta ocorrência em vários tipos de rochas e suas origens diversas.(30) A composição química deste mineral é Fe_2O_3 , com uma fração mássica em ferro de 69,9% e uma fração mássica em oxigênio de 30,1%.(31)

...

Table 2.1: Main morphologies of hematite.(14)

Tipo	Características	Forma Textura	Ilustração Esquemática
Hematita Microcristalina	<ul style="list-style-type: none"> ▷ Cristais muito pequenos, < 0.01 mm. ▷ Textura porosa. ▷ Contatos pouco desenvolvidos. 		
Magnetita	<ul style="list-style-type: none"> ▷ Cristais euédricos isolados ou em agregados. ▷ Cristais compactos. 		
Martita	<ul style="list-style-type: none"> ▷ Hematita com hábito de magnetita. ▷ Oxidação segundo os planos cristalográficos da magnetita. ▷ Geralmente porosa. 		
Hematita Lobular	<ul style="list-style-type: none"> ▷ Formatos irregulares inequidimensionais. ▷ Contatos irregulares, geralmente imbricados. 		
Hematita Granular	<ul style="list-style-type: none"> ▷ Formatos regulares equidimensionais. ▷ Contatos retilíneos e junções triplíceis. ▷ Cristais compactos. 		
Hematita Lamelar	<ul style="list-style-type: none"> ▷ Cristais inequidimensionais, hábito tabular. ▷ Contato retilíneo. ▷ Cristais compactos. 		
Hidróxidos de Fe (Goethita-Limonita)	<ul style="list-style-type: none"> ▷ Material cripto-cristalino. ▷ Estrutura colorforme, hábito botrioidal. ▷ Textura porosa. 		

3

Conclusions

Um sistema de microscopia digital com reconhecimento e classificação automática dos cristais de hematita em minérios de ferro foi desenvolvido.

O método utiliza operações tradicionais de processamento digital de imagens e propõe uma segmentação automática de cristais baseada no cálculo da distância espectral, a fim de controlar ...

É fundamental também comentar que ...

Assim, como uma proposta para trabalho futuro, pode-se buscar combinar os dois enfoques...

Bibliography

- [1] LODISH, H.; BERK, A.; MATSUDAIRA, P.; KAISER, C.; KRIEGER, M. ; SCOTT, M.. **Molecular Cell Biology**, volumen 60. New York, 5th edition, 1990.
- [2] LEHNINGER, A.; NELSON, D. ; COX, M.. **Principles of Biochemistry**, volumen 17. New York, 4th edition, 2005.
- [3] CREIGHTON, T. E.. **Protein folding**. *Biochem. J.*, 270(1):1–16, 1990.
- [4] LESK, A.. **Introduction to Bioinformatics**, volumen I. New York, 1st edition, 2002.
- [5] BRANDEN, C.; TOOZE, J.. **Introduction to Protein Structure**. Garland Publishing Inc, New York, 2nd edition, 1998.
- [6] BAXEVANIS, A.; OUELLETTE, B.. **Bioinformatics: a practical guide to the analysis of genes and proteins**. 2001.
- [7] SERINGHAUS, M.. **Developing Bioinformatics Computer Skills**, volumen 75. 2002.
- [8] GÜNTERT, P.. **Automated NMR structure calculation with CYANA**. *Methods in molecular biology* (Clifton, N.J.), 278:353–78, 2004.
- [9] BUJNICKI, J. M.. **Protein-structure prediction by recombination of fragments**. *Chembiochem : a European journal of chemical biology*, 7(1):19–27, jan 2006.
- [10] MOULT, J.. **A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction**, jun 2005.
- [11] OSGUTHORPE, D.. **Ab initio protein folding**. *Current Opinion in Structural Biology*, 10(2):146–152, apr 2000.
- [12] TRAMONTANO, A.. **Protein Structure Prediction**. John Wiley and Sons, 1st edition, 2006.

- [13] FLOUDAS, C.; FUNG, H.; MCALLISTER, S.; MÖNNIGMANN, M. ; RAJ-GARIA, R.. **Advances in protein structure prediction and de novo protein design: A review.** Chemical Engineering Science, 61(3):966–988, feb 2006.
- [14] BOWIE, J. U.; LUTHY, R. ; EISENBERG, D.. **A method to identify protein sequences that fold into a known three- dimensional structure.** Science, 253(5016):164–170, 1991.
- [15] **A new approach to protein fold recognition.** Nature, 358(6381):86–89, jul 1992.
- [16] BRYANT, S. H.; ALTSCHUL, S. F.. **Statistics of sequence-structure threading.** Current Opinion in Structural Biology, 5(2):236–244, 1995.
- [17] TATTERSALL, S. F.; BENSON, M. K.; HUNTER, D.; MANSELL, A.; PRIDE, N. B. ; FLETCHER, C. M.. **The use of tests of peripheral lung function for predicting future disability from airflow obstruction in middle-aged smokers.** The American review of respiratory disease, 118(6):1035–50, dec 1978.
- [18] MARTÍ-RENO, M. A.; STUART, A. C.; FISER, A.; SÁNCHEZ, R.; MELO, F. ; ŠALI, A.. **Comparative Protein Structure Modeling of Genes and Genomes.** Annual Review of Biophysics and Biomolecular Structure, 29(1):291–325, jun 2000.
- [19] SÁNCHEZ, R.; ŠALI, A.. **Advances in comparative protein-structure modelling.** Current Opinion in Structural Biology, 7(2):206–214, apr 1997.
- [20] **Protein Structure Prediction Using Rosetta.** Methods in Enzymology, 383(2003):66–93, 2004.
- [21] **LINUS: A hierarchic procedure to predict the fold of a protein.** Proteins: Structure, Function, and Bioinformatics, 22(2):81–99, jun 1995.
- [22] KOLINSKI, A.. **Protein modeling and structure prediction with a reduced representation.** Acta Biochimica Polonica, 51(2):349–371, 2004.
- [23] **Ranking predicted protein structures with support vector regression.** Proteins: Structure, Function and Genetics, 71(3):1175–1182, nov 2008.

- [24] **Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11.** *Proteins: Structure, Function and Bioinformatics*, 2014(May 2014):349–369, 2015.
- [25] JING, X.; WANG, K.; LU, R. ; DONG, Q.. **Sorting protein decoys by machine- learning-to-rank.** *Nature Publishing Group*, (April):1–11, 2016.
- [26] BRÜSCHWEILER, R.. **Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation.** *Proteins*, 50(1):26–34, jan 2003.
- [27] ZEMLA, A.. **LGA: A method for finding 3D similarities in protein structures.** *Nucleic acids research*, 31(13):3370–4, jul 2003.
- [28] ZHANG, Y.; SKOLNICK, J.. **Scoring function for automated assessment of protein structure template quality.** *Proteins*, 57(4):702–10, dec 2004.
- [29] SIEW, N.; ELOFSSON, A.; RYCHLEWSKI, L. ; FISCHER, D.. **MaxSub: an automated measure for the assessment of protein structure prediction quality.** *Bioinformatics (Oxford, England)*, 16(9):776–785, 2000.
- [30] LAZARIDIS, T.; KARPLUS, M.. **Discrimination of the native from misfolded protein models with an energy function including implicit solvation.** *Journal of molecular biology*, 288(3):477–87, may 1999.
- [31] **Free energy determinants of tertiary structure and the evaluation of protein models.** *Protein science : a publication of the Protein Society*, 9(11):2181–91, nov 2000.
- [32] LU, M.; DOUSIS, A. D. ; MA, J.. **OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing.** *Journal of molecular biology*, 376(1):288–301, feb 2008.
- [33] ZHOU, H.; ZHOU, Y.. **Distance scaled, finite ideal gas reference state improves structure derived potentials of mean force for structure selection and stability prediction.** *Protein science*, 11:2714–2726, 2002.
- [34] **Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust.** *Proteins*, 77 Suppl 9(SUPPL. 9):173–80, 2009.

- [35] WALLNER, B.; ELOFSSON, A.. Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins*, 69 Suppl 8(S8):184–93, 2007.
- [36] GINALSKI, K.; ELOFSSON, A.; FISCHER, D. ; RYCHLEWSKI, L.. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics (Oxford, England)*, 19(8):1015–8, may 2003.
- [37] WANG, Z.; TEGGE, A. N. ; CHENG, J.. Evaluating the absolute quality of a single protein model using structural features and support vector machines. *Proteins*, 75(3):638–47, may 2009.
- [38] RANA, P. S.; SHARMA, H.; BHATTACHARYA, M. ; SHUKLA, A.. Quality assessment of modeled protein structure using physicochemical properties. *Journal of Bioinformatics and Computational Biology*, 13(02):1550005, 2015.
- [39] FARAGGI, E.; KLOCZKOWSKI, A.. A global machine learning based scoring function for protein structure prediction. *Proteins*, 82(5):752–9, 2014.
- [40] A machine learning-based method for protein global model quality assessment. *International Journal of General Systems*, 40(4):417–425, 2011.
- [41] GHOSH, S.; VISHVESHWARA, S.. Ranking the quality of protein structure models using sidechain based network properties. *F1000Research*, 17(MAY):1–9, 2014.

A

Published paper

The following paper was published ...