SHREC 2021: Retrieval and classification of protein surfaces equipped with physical and chemical properties

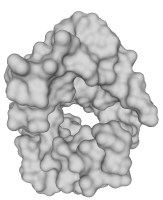
Andrea Raffo, Ulderico Fugacci, Silvia Biasotti, Walter Rocchia, Yonghuai Liu, Ekpo Otu, Reyer Zwiggelaar, David Hunter, Evangelia I. Zacharaki, Eleftheria Psatha, Dimitrios Laskos Gerasimos Arvanitis, Konstantinos Moustakas, Tunde Aderinwale, Charles Christoffer, Woong-Hee Shin, Daisuke Kihara, Andrea Giachetti, Huu-Nghia Nguyen, Tuan-Duy Nguyen, Vinh-Thuyen Nguyen-Truong, Danh Le-Thanh, Hai-Dang Nguyen and Minh-Triet Tran

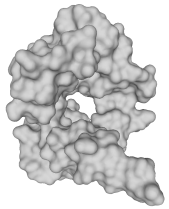
3D Object Retrieval 2021 (3DOR'21)
POR FSE Liguria 2014-2020

Motivation

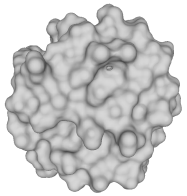
Proteins are generally thought to adopt unique structures determined by their amino acid sequences, which are crucial for their functions.

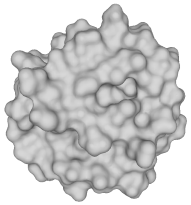
However, proteins are not strictly static objects, but rather populate **ensembles of conformations**.





CALMODULIN-1





THIOREDOXIN

Recognising a protein from an ensemble of geometries can be assumed to mean capturing the **features that are unique** to it. It is preliminary to the definition of a geometry-based notion of similarity, and, subsequently, complementarity, between proteins.

Computers & Graphics 99 (2021) 1–21

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on 3DOR 2021

SHREC 2021 Track: Retrieval and classification of protein surfaces equipped with physical and chemical properties

Andrea Raffo[a,1,*], Ulderico Fugacci[a,1], Silvia Biasotti[a,1], Walter Rocchia[b,1], Yonghuai Liu[c], Ekpo Otu[d], Reyer Zwiggelaar[d], David Hunter[d], Evangelia I. Zacharaki[e], Eleftheria Psatha[e], Dimitrios Laskos[e], Gerasimos Arvanitis[e], Konstantinos Moustakas[e], Tunde Aderinwale[f], Charles Christoffer[f], Woong-Hee Shin[g], Daisuke Kihara[h,f], Andrea Giachetti[i], Huu-Nghia Nguyen[j,l], Tuan-Duy Nguyen[j,l], Vinh-Thuyen Nguyen-Truong[j,l], Danh Le-Thanh[j,l], Hai-Dang Nguyen[j,l], Minh-Triet Tran[j,k,l]

[a] Istituto di Matematica Applicata e Tecnologie Informatiche "E. Magenes", Consiglio Nazionale delle Ricerche, Genova, Italy
[b] CONCEPT Lab, Istituto Italiano di Tecnologia, Genova, Italy
[c] Department of Computer Science, Edge Hill University, Ormskirk, UK
[d] Department of Computer Science, Aberystwyth University, Aberystwyth, UK
[e] Department of Electrical and Computer Engineering, University of Patras, Patras, Greece
[f] Department of Computer Science, Purdue University, West Lafayette, USA
[g] Department of Chemical Science Education, Sunchon National University, Suncheon, Republic of Korea
[h] Department of Biological Sciences, Purdue University, West Lafayette, USA
[i] Department of Computer Science, University of Verona, Verona, Italy
[j] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[k] John von Neumann Institute, VNU-HCM, Ho Chi Minh City, Vietnam
[l] Vietnam National University, Ho Chi Minh City, Vietnam

ARTICLE INFO

Article history:
Received 1 April 2021
Revised 28 May 2021
Accepted 21 June 2021
Available online 25 June 2021

Keywords:
SHREC
Protein surfaces
Protein retrieval
Protein classification
3D Shape analysis
3D Shape descriptor

ABSTRACT

This paper presents the methods that have participated in the SHREC 2021 contest on retrieval and classification of protein surfaces on the basis of their geometry and physicochemical properties. The goal of the contest is to assess the capability of different computational approaches to identify different conformations of the same protein, or the presence of common sub-parts, starting from a set of molecular surfaces. We addressed two problems: defining the similarity solely based on the surface geometry or with the inclusion of physicochemical information, such as electrostatic potential, amino acid hydrophobicity, and the presence of hydrogen bond donors and acceptors. Retrieval and classification performances, with respect to the single protein or the existence of common sub-sequences, are analysed according to a number of information retrieval indicators.

1. Introduction

Automatically identifying the different conformations of a given set of proteins, as well as their interaction with other molecules, is crucial in structural bioinformatics. The well-established shape-function paradigm for proteins [1] states that a protein of a given sequence has one main privileged conformation, which is crucial for its function. However, every protein during its time evolution explores a much larger part of the conformational space. The most stable conformations visited by the protein can be experimentally captured by the NMR technique; this is because the hydrogen atoms are already included in the atomic model, thus giving less ambiguities in the charge assignment.

Recognising a protein from an ensemble of geometries corresponding to the different conformations it can assume means capturing the features that are unique to it and is a fundamental step from the structural bioinformatics viewpoint. It is preliminary to the definition of a geometry-based notion of similarity, and, subsequently, complementarity, between proteins. From the application standpoint, the identification of characteristic features can point

* Corresponding author.
E-mail addresses: andrea.raffo@ge.imati.cnr.it (A. Raffo), ulderico.fugacci@ge.imati.cnr.it (U. Fugacci), silvia.biasotti@ge.imati.cnr.it (S. Biasotti), walter.rocchia@iit.it (W. Rocchia).
[1] Track organizer.

https://doi.org/10.1016/j.cag.2021.06.010
0097-8493/

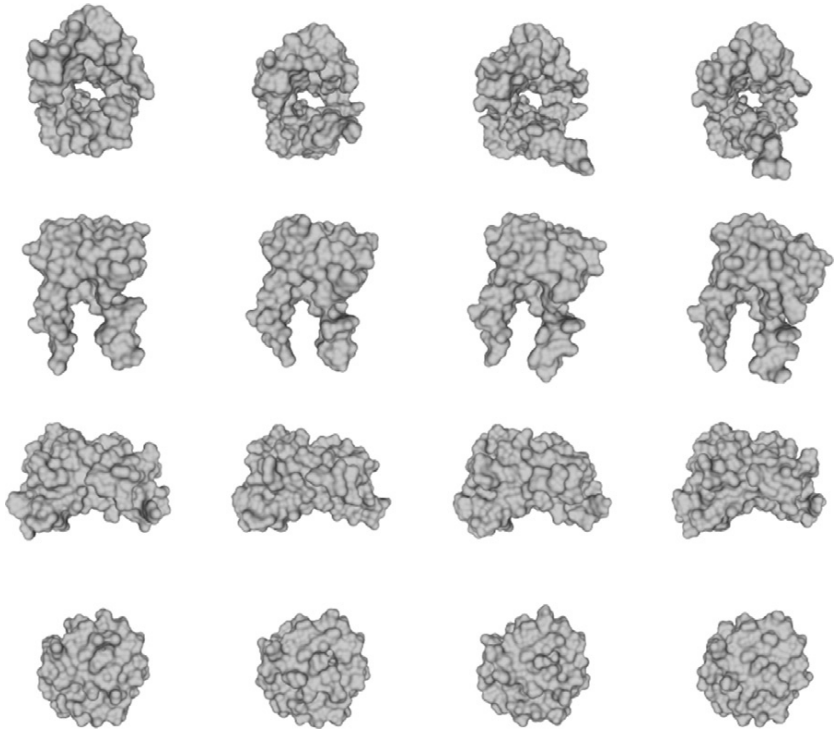
Unione Europea
Fondo europeo sociale

The benchmark

The dataset

- 5,128 **triangle meshes** of protein surfaces:
  - ★ Available in the OFF[1] format
  - ★ Approximate the Solvent Excluded Surface (SES)
  - ★ Computed by NanoShaper [1]
  - ★ Pre-split: 70% training set, 30% test set

- For each surface, **three physicochemical proteins** were approximated at its vertices:
  - (a) Electrostatic potential ← Delphi [2,3]
  - (b) Hydrophobicity
  - (c) Hydrogen bond donors and acceptors } ← Own routines



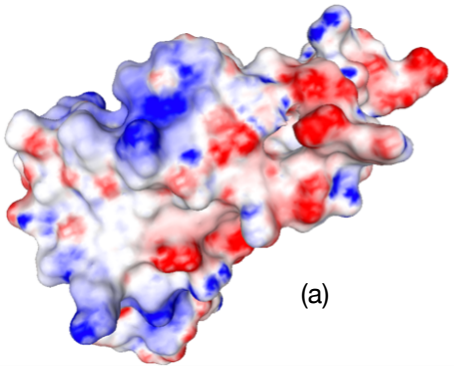
[1] https://segeval.cs.princeton.edu/public/off_format.html

The dataset

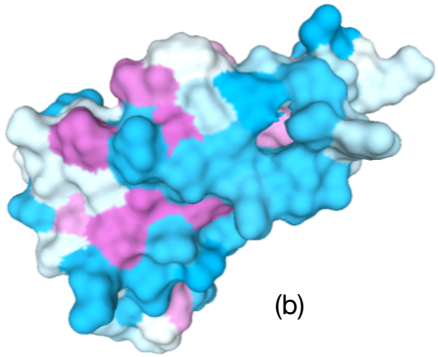
- 5,128 **triangle meshes** of protein surfaces:
  - ★ Available in the OFF[1] format
  - ★ Approximate the Solvent Excluded Surface (SES).
  - ★ Computed by NanoShaper [1]
  - ★ Pre-split: 70% training set, 30% test set

- For each surface, **three physicochemical proteins** were approximated at its vertices:
  - (a) Electrostatic potential ← Delphi [2,3]
  - (b) Hydrophobicity
  - (c) Hydrogen bond donors and acceptors

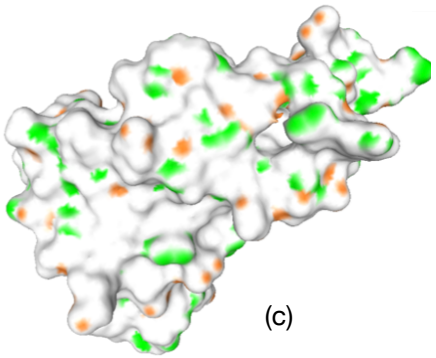
  } ← Own routines



(a)



(b)



(c)

[1] https://segeval.cs.princeton.edu/public/off_format.html

The ground truth

Evaluation of the results will be based on two ground truths:

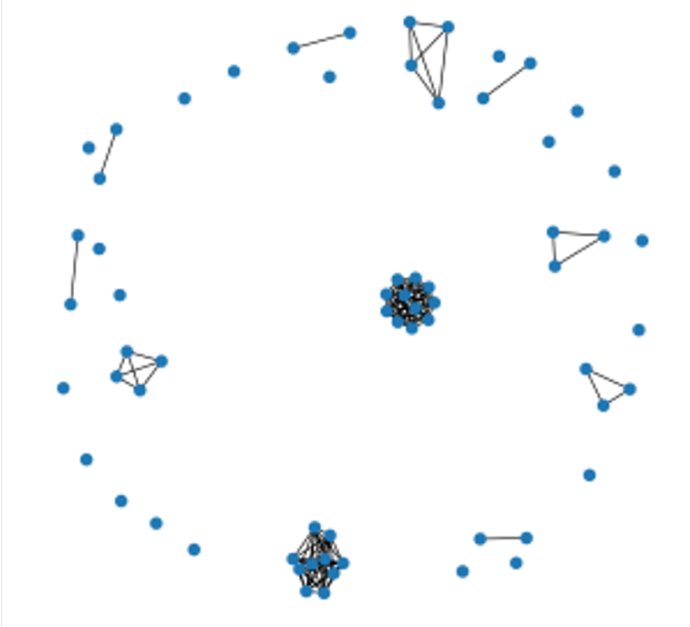
- A 2-level classification of the dataset (**PDB-based classification**)
- A 4-level classification of the dataset (**BLAST-based classification**)

PDB-based classification Two protein surfaces are conformations of the same protein if and only if they refer to the same PDB code.

BLAST-based classification We introduce a less strict classification on the basis of the concept of sequence similarity λ:

- Extremely similar $\longrightarrow$ $95\,\% \leq \lambda$ AND at least 50 aligned residuals
- Highly related $\longrightarrow$ $35\,\% \leq \lambda < 95\,\%$ AND at least 50 aligned residuals
- Similar $\longrightarrow$ $28\,\% \leq \lambda < 35\,\%$ AND at least 50 aligned residuals
- Dissimilar $\longrightarrow$ $\lambda < 28\,\%$ OR less than 50 aligned residuals

Obtained classes are refined for ensuring transitivity.



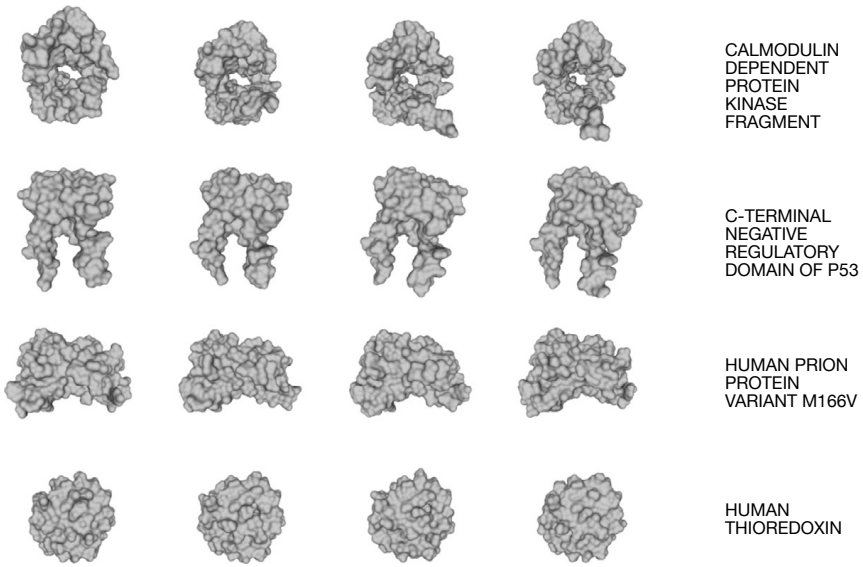
Nodes ↕ PDB-based classes

Connected components ↕ BLAST-based classes

The accuracy measures

- **Retrieval measures**:
  - ★ Precision-recall curves and mean Average Precision (mAP)
  - ★ First and Second Tiers (1T, 2T)
  - ★ Normalized Discounted Cumulated Gain (NDCG)
  - ★ Average Dynamic Recall (ADR)

- **Classification measures**:
  - ★ True Positive and Negative Rates (TPR, TNR)
  - ★ Positive and Negative Predicted Values (PPV, NPV)
  - ★ Accuracy (ACC)
  - ★ F1-score



Proposed methods

Proposed methods

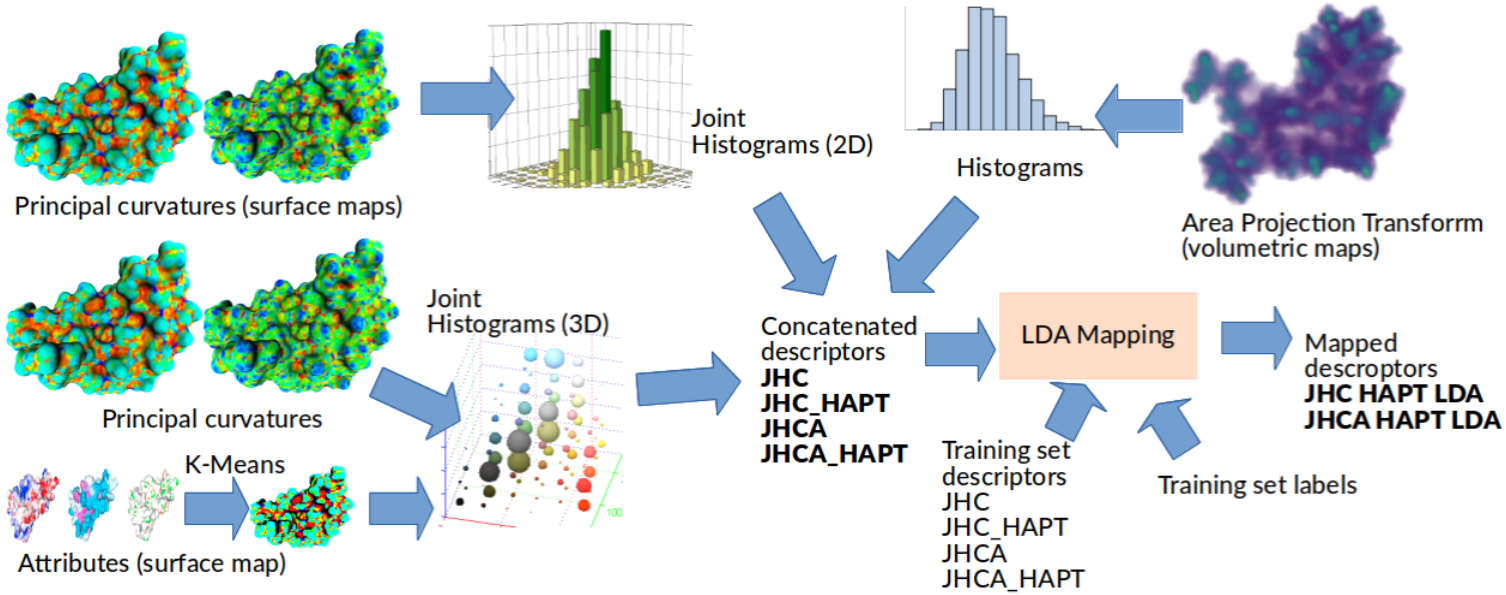
- **P1: “Joint histograms of curvatures, local properties and area projection transform”** by Andrea Giachetti
- **P2: “3D Zernike descriptors”** by Tunde Aderinwale, Charles Christoffer, Woong-Hee Shin, and Daisuke Kihara
- **P3: “Hybrid Augmented Point Pair Signatures and Histogram of Processed Physicochemical Properties of Protein molecules ”** by Yonghuai Liu, Ekpo Otu, Reyer Zwiggelaar, and David Hunter
- **P4: “Global and Local Feature fit”** by Evangelia I. Zacharaki, Eleftheria Psatha, Dimitrios Laskos, Gerasimos Arvanitis, and Konstantinos Moustakas
- **P5: “Message-Passing Graph Convolutional Neural Networks (MPGCNNs) and PointNet ”** by Huu-Nghia Nguyen, Tuan-Duy Nguyen, Vinh-Thuyen Nguyen-Truong, Danh Le-Thanh, Hai- Dang Nguyen, and Minh-Triet Tran

Two tasks were proposed to the participants, with up to three runs per task:

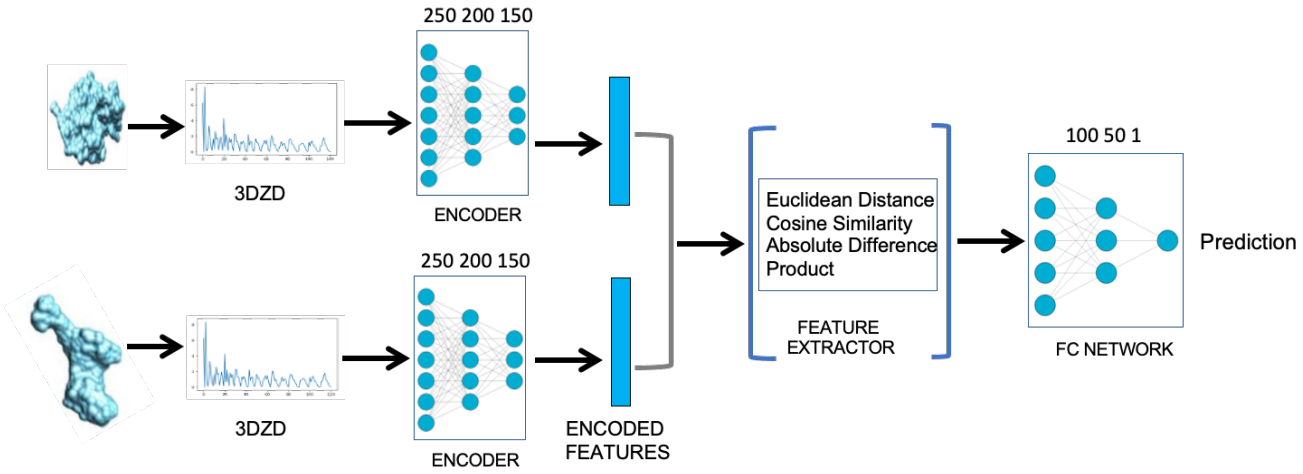
Task A: only the OFF files of the models are to be considered (i.e., only the geometry).
Task B: in addition to the geometry, physicochemical properties are to be considered.

P1: “Joint histograms of curvatures, local properties and area projection transform”

- Use of surface properties (**curvatures+attributes**) combined with joint histograms
- Combined with **volumetric** (symmetry) **features**: Area Projection Transform (Giachetti & Lovato 2012)
- Test of supervised **dimensionality reduction** (Linear Discriminant Analysis)
  - ★ Using training set labels
  - ★ Not effective (different classes)



P2: “3D Zernike descriptors”



	NN	1T	2T
Validation Results			
SHAPE: EXTRACTOR	0.854	0.788	0.442
SHAPE: AVG of (EXTRACTOR&End2End)	*0.862*	*0.785*	*0.442*
SHAPE: AVG of (E2E & EUCLID. of 3DZD)	*0.894*	*0.822*	*0.447*
SHAPE+Phys: EXTRACTOR	0.889	0.858	0.456
SHAPE+Phys: AVG of (EXTRACTOR & EUCLID)	0.899	0.861	0.457
SHAPE+Phys: AVG of (EXTRACTOR&E2E)	*0.894*	*0.870*	*0.459*

- Total number of (training) surfaces provided: 3,585
- Training-validation set split : 80%/20%
- Total number of validation proteins: 717
- Out of 717*716/2 pairs, 10,436 pairs were used for validation

- Shape only: 3ZD for protein surface only
- Shape + Phys: 3D3Z for protein surface and physicochemical properties

P3: “Hybrid Augmented Point Pair Signatures and Histogram of Processed Physicochemical Properties of Protein molecules ”

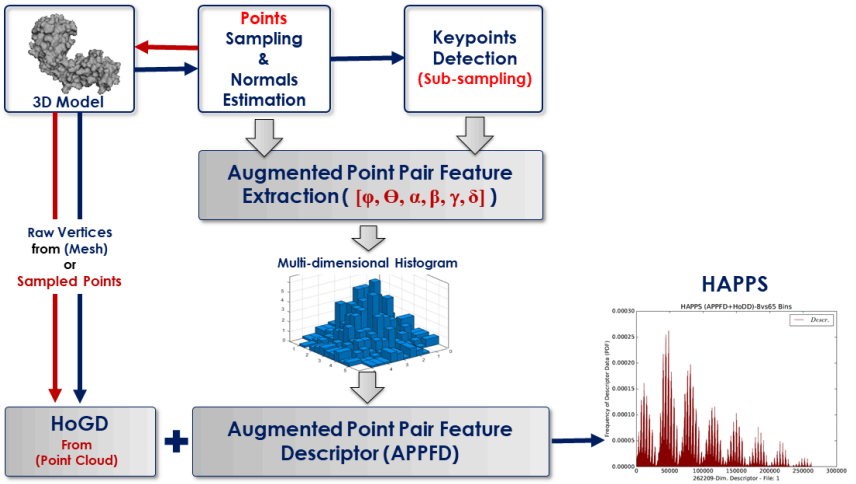
Two separate retrieval strategies for the two different tasks

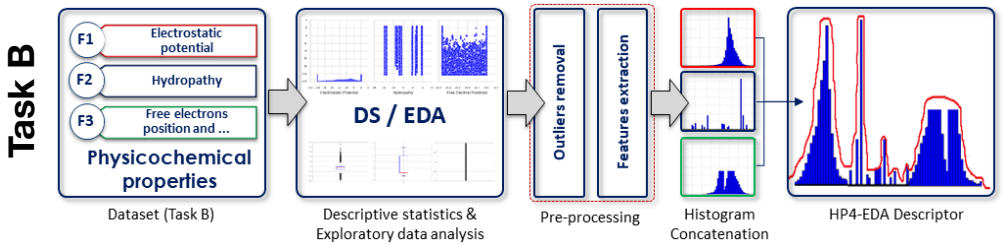
Task A: Hybrid Augmented Point Pair Signature (**HAPPS**)

- A 3D geometric shape descriptor combining a collection of normalized vectors between the point of the surface and its centroid with the local geometry around each point

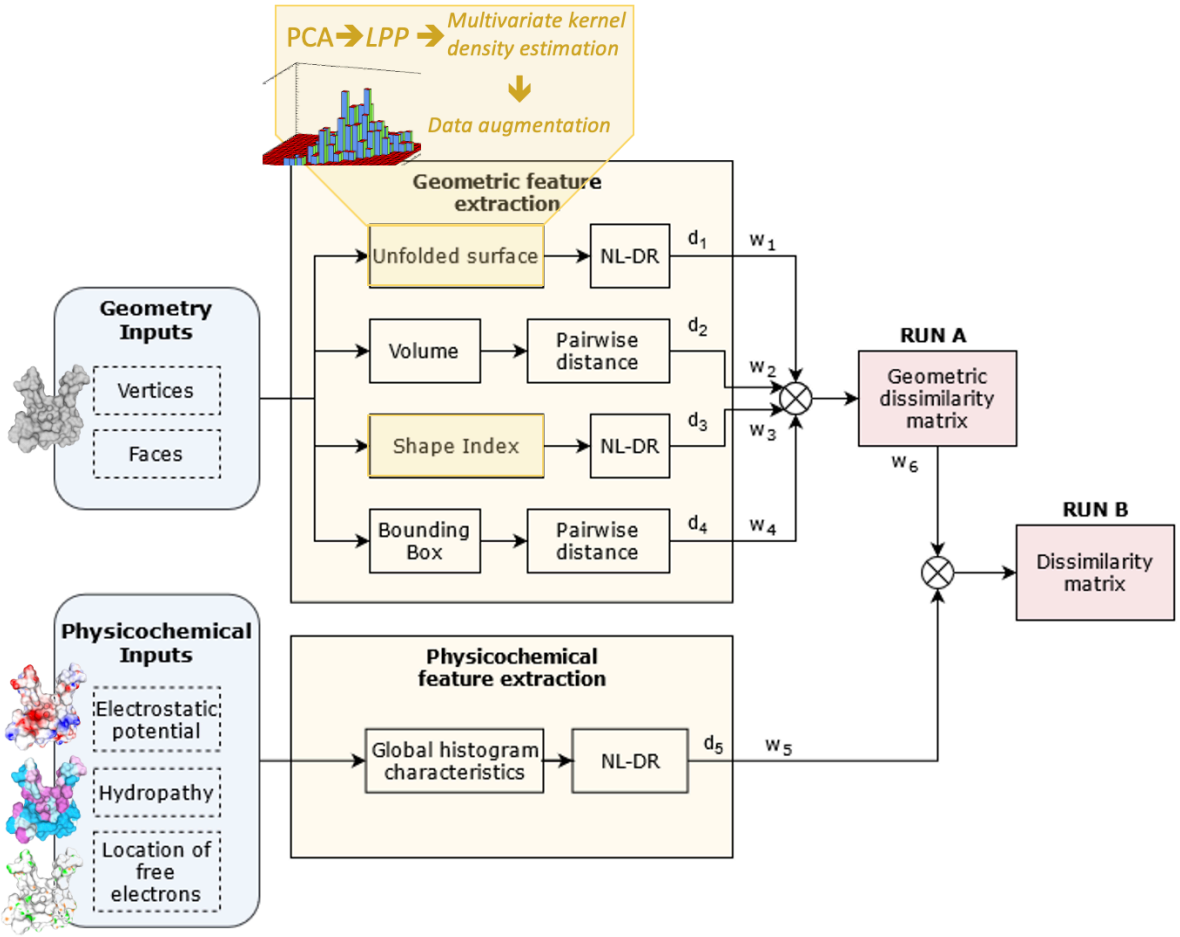
Task B: Histogram of Processed Physicochemical Properties of Protein following an Exploratory Data Analysis (**HP4-EDA**)

- Strategy based on a descriptive statistics (DS) of the 3D physicochemical variables, following an exploratory data analysis (EDA) of each of them.





P4: “Global and Local Feature fit”



NL-DR : Non-Linear Dimensionality Reduction ➔ tSNE

Shape Index $= \frac{2}{\pi} tan^{-1} \frac{k_1 + k_2}{k_1 - k_2}$

where k_1, k_2 ($k_1 \geq k_2$) are the principal curvature values

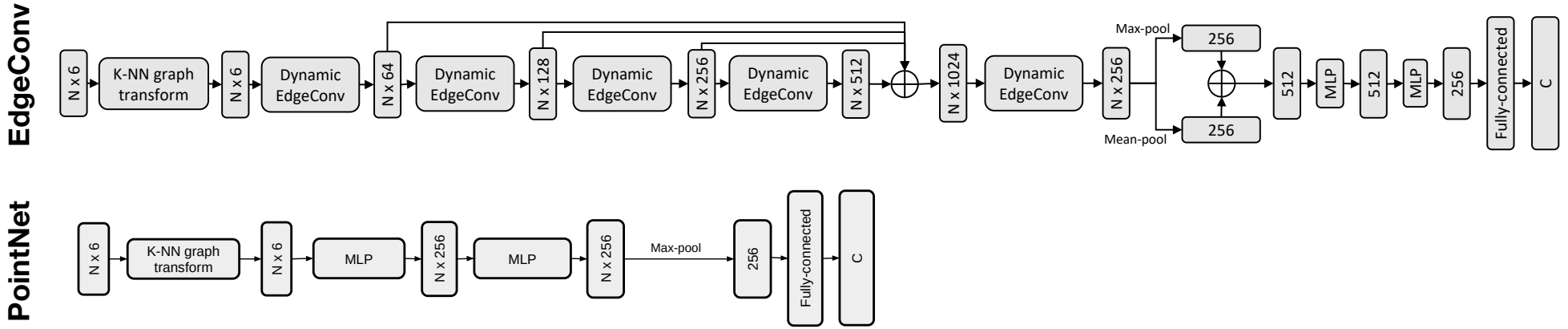
For each of the 3 physicochemical properties:
global histogram characteristics (mean intensity, standard deviation, mode of histogram, kurtosis, skewness, and energy)

P5: “Message-Passing Graph Convolutional Neural Networks (MPGCNNs) and PointNet ”

Strategy based on the use **graph neural networks** (**GNNs**): a deep learning based methods operating on generalized graph domains rather than on Euclidean ones

Two main network architectures have been adopted:

- **EdgeConv.** The module that performs the graph message-passing function is a dynamic variant of edge convolution
- **PointNet.** Two message-passing modules each containing a MLP block that uses ReLU as activation function



Evaluation and results

Complete analysis in the paper!

Results — PDB-based (2-level) classification

From the classification measures:

- All methods have TNR higher than TPR ⟶ more reliable in finding true negatives
- All methods have NPV higher than TPV ⟶ more reliable in reporting negatives
- Accuracy is high (always above 95%), but F1-score is a better indicator (classes are unbalanced!)
- Physicochemical properties can bring additional information, but it does NOT always mean an improvement

	Geometry						Geometry and physicochemical properties						
	method	TPR	TNR	PPV	NPV	ACC	F1	method	TPR	TNR	PPV	NPV	ACC	F1
P1: Joint histograms	run 1	0.8373	0.9967	0.8401	0.9973	0.9941	0.8354	run 1	0.9825	0.9997	0.9860	0.9997	0.9994	0.9832
	run 2	**0.9475**	**0.9991**	**0.9489**	**0.9992**	**0.9983**	**0.9467**	run 2	**0.9890**	**0.9999**	**0.9921**	**0.9999**	**0.9998**	**0.9893**
	run 3	0.9274	0.9990	0.9304	0.9988	0.9974	0.9239	run 3	0.5839	0.9885	0.6086	0.9912	0.9807	0.5727
P2: 3D Zernike + ML	run 1	0.9145	0.9979	0.9159	0.9984	0.9965	0.9119	run 1	0.9514	0.9989	0.9525	0.9992	0.9981	0.9504
	run 2	0.8944	0.9975	0.8977	0.9976	0.9954	0.8931	run 2	0.9469	0.9989	0.9470	0.9991	0.9981	0.9460
	run 3	**0.9242**	**0.9983**	**0.9253**	**0.9985**	**0.9969**	**0.9222**	run 3	**0.9793**	**0.9995**	**0.9819**	**0.9996**	**0.9991**	**0.9791**
P3: Point pair signatures	run 1	0.9196	**0.9985**	0.9205	0.9986	**0.9971**	0.9169	run 1	0.9015	0.9977	0.9037	0.9979	0.9957	0.9007
	run 2	**0.9300**	0.9982	**0.9333**	**0.9988**	**0.9971**	**0.9276**	run 2	**0.9216**	**0.9985**	**0.9238**	**0.9987**	**0.9974**	**0.9207**
	run 3	0.9222	0.9982	0.9258	0.9986	0.9969	0.9199	run 3	0.9015	0.9979	0.9005	0.9985	0.9966	0.8987
P4: Global & Local Fit	run 1	**0.9274**	**0.9983**	**0.9284**	**0.9987**	**0.9971**	**0.9262**	run 1	**0.9410**	0.9987	**0.9424**	**0.9990**	**0.9978**	0.9406
	run 2	0.9268	**0.9983**	0.9277	**0.9987**	**0.9971**	0.9252	run 2	**0.9410**	**0.9988**	0.9422	**0.9990**	**0.9978**	**0.9409**
	run 3	0.9067	0.9979	0.9059	0.9983	0.9963	0.9046	run 3	0.9326	0.9984	0.9336	0.9988	0.9974	0.9323
P5: Graph CNN	run 1	**0.7537**	**0.9944**	**0.7539**	**0.9943**	**0.9892**	**0.7507**	run 1	**0.7187**	**0.9937**	**0.7215**	**0.9940**	**0.9886**	**0.7160**
	run 2	0.4362	0.9870	0.4412	0.9873	0.9754	0.4328							
	run 3	0.7123	0.9927	0.7109	0.9930	0.9868	0.7044							

Complete analysis in the paper!

Results — BLAST-based (4-level) classification

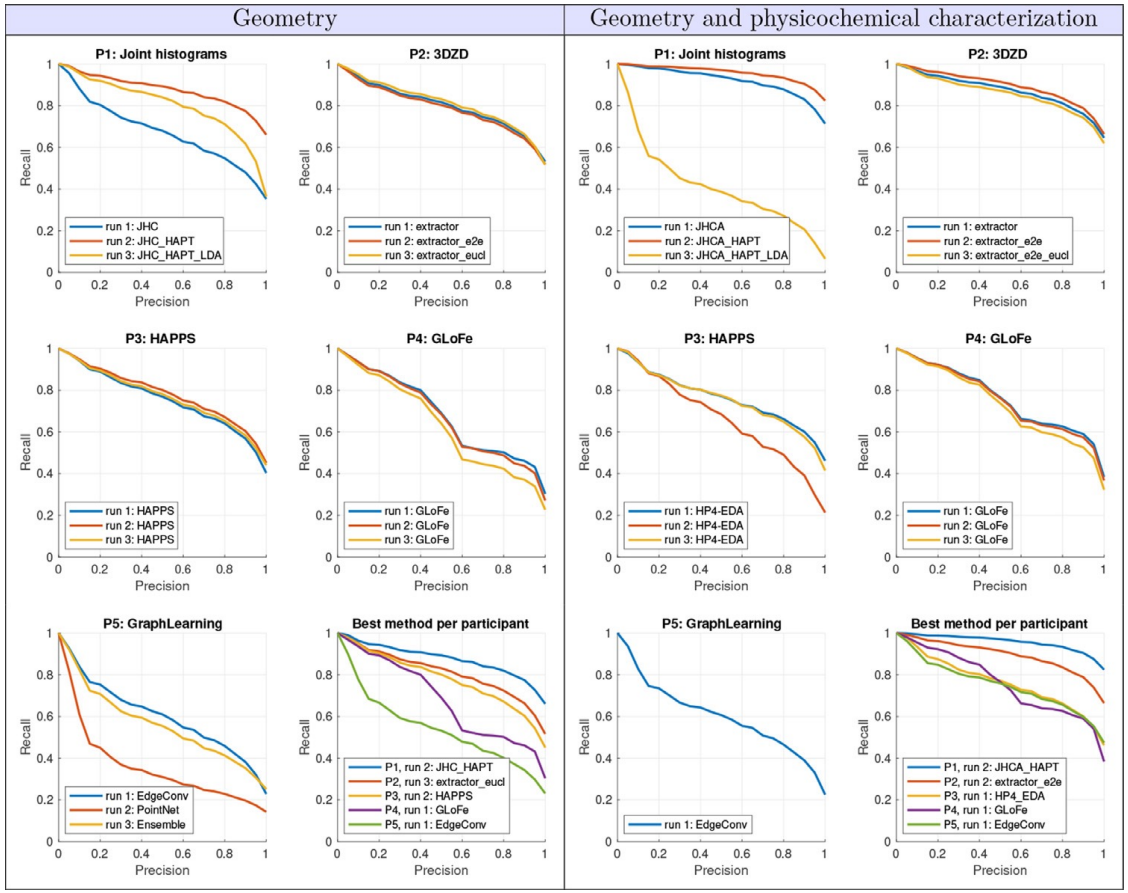
From the classification measures:

- As expected, decreasing the number of classes improves the classification performance
- The improvement takes place despite only PDB-based classification being available to the participants (maybe not surprising, as the number of classes is reduced)

	Geometry							Geometry and physicochemical properties						
	method	TPR	TNR	PPV	NPV	ACC	F1	method	TPR	TNR	PPV	NPV	ACC	F1
P1: Joint histograms	run 1	0.9086	0.9949	0.9082	0.9959	0.9917	0.9069	run 1	0.9961	0.9998	0.9962	**1.0000**	0.9997	0.9960
	run 2	**0.9890**	**0.9996**	**0.9891**	0.9996	**0.9993**	**0.9888**	run 2	**0.9981**	**1.0000**	**0.9981**	**1.0000**	**1.0000**	**0.9980**
	run 3	0.9844	**0.9996**	0.9869	**0.9997**	**0.9993**	0.9840	run 3	0.8529	0.9904	0.8562	0.9931	0.9854	0.8452
P2: 3D Zernike + ML	run 1	0.9760	**0.9992**	0.9766	**0.9993**	**0.9985**	0.9758	run 1	0.9929	0.9997	0.9930	0.9999	0.9996	0.9928
	run 2	0.9728	0.9991	0.9746	0.9991	0.9983	0.9723	run 2	0.9909	0.9998	0.9909	0.9999	0.9997	0.9908
	run 3	**0.9767**	0.9985	**0.9770**	0.9989	0.9977	**0.9764**	run 3	**0.9987**	**0.9999**	**0.9987**	**1.0000**	**0.9999**	**0.9987**
P3: Point pair signatures	run 1	0.9689	0.9981	0.9690	0.9985	0.9970	0.9680	run 1	**0.9942**	0.9996	**0.9943**	**0.9999**	0.9995	**0.9942**
	run 2	**0.9747**	0.9980	**0.9754**	**0.9989**	0.9972	**0.9740**	run 2	0.9916	**0.9997**	0.9921	0.9998	**0.9996**	0.9916
	run 3	0.9721	**0.9987**	0.9738	0.9985	**0.9976**	0.9716	run 3	0.9806	0.9981	0.9809	0.9994	0.9980	0.9803
P4: Global & Local Fit	run 1	**0.9799**	**0.9987**	**0.9805**	**0.9991**	**0.9980**	**0.9798**	run 1	**0.9903**	**0.9995**	**0.9909**	**0.9996**	**0.9992**	**0.9904**
	run 2	0.9793	**0.9987**	0.9801	0.9990	0.9979	0.9791	run 2	**0.9903**	**0.9995**	0.9908	**0.9996**	0.9991	**0.9904**
	run 3	0.9734	0.9984	0.9738	0.9987	0.9974	0.9732	run 3	0.9870	0.9993	0.9876	0.9994	0.9988	0.9871
P5: Graph CNN	run 1	**0.9209**	**0.9953**	**0.9209**	**0.9958**	**0.9923**	**0.9197**	run 1	**0.9501**	**0.9975**	**0.9506**	**0.9988**	**0.9968**	**0.9491**
	run 2	0.7168	0.9852	0.7201	0.9853	0.9734	0.7156							
	run 3	0.9047	0.9944	0.9031	0.9953	0.9910	0.9019							

Complete analysis in the paper!

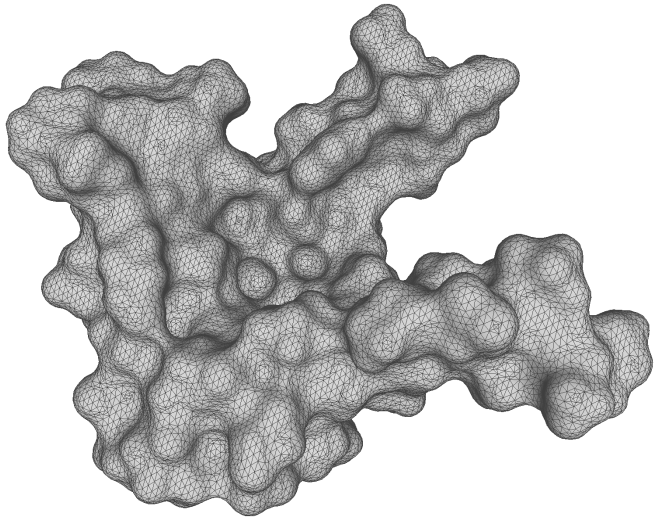
Results— PDB- and BLAST-based classifications



Unexpectedly, the two ground truths have much in common:

- Methods are good but not optimal: highest mAP score is 0.87 (only geometry) and 0.96 (geometry + chemistry) —> e.g., P2
- The strong heterogeneity of the class size has influenced prediction accuracy, especially for learning-based methods —> e.g., P5
- Physicochemical properties can provide additional information, but one must be careful how to use it —> e.g., P1, run 3
- Deep learning does NOT guarantee a better performance (risk to overfit data!) —> e.g., P5

Conclusions



- W.r.t. previous contests on protein retrieval, we have taken into account physicochemical properties, provided the participants of a training set and a test set, and proposed different ground truths
- The number of registered teams (8) and of actual participants (5) shows the interest of the community to the problem, despite some difficulties
- The methods present a satisfactory variety in terms of the paradigms nowadays popular

We are underway to apply for the replicability stamp. The benchmark will be available at: https://github.com/rea1991/SHREC2021

Thanks for your attention!

You can reach me at andrea.raffo@ge.imati.cnr.it

References

[1] S. Decherchi and W. Rocchia. A general and robust ray-casting-based algorithm for triangulating surfaces at the nanoscale. *PLoS ONE,* vol.8, no. 4, pp. 1—15, 2013.

[2] W. Rocchia, E. Alexov, and B. Honig. “Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions”. *The Journal of Physical Chemistry B*, vol. 105, no. 28, pp. 6507–6514, 2001.

[3] L. Wang, Z. Zhang, W. Rocchia, and E. Alexov, “Using delphi capabilities to mimic protein’s conformational reorganization with amino acid specific dielectric constants,” *Communications in Computational Physics*, vol. 13, no. 1, pp. 13–30, 2013.

[4] S. Fortunato, “Community detection in graphs,” *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.