Machine learning-based prediction of coil conversion to β-strand in dimer formation

Gabriel Flores[1]; Jean-Christophe Nebel[2]

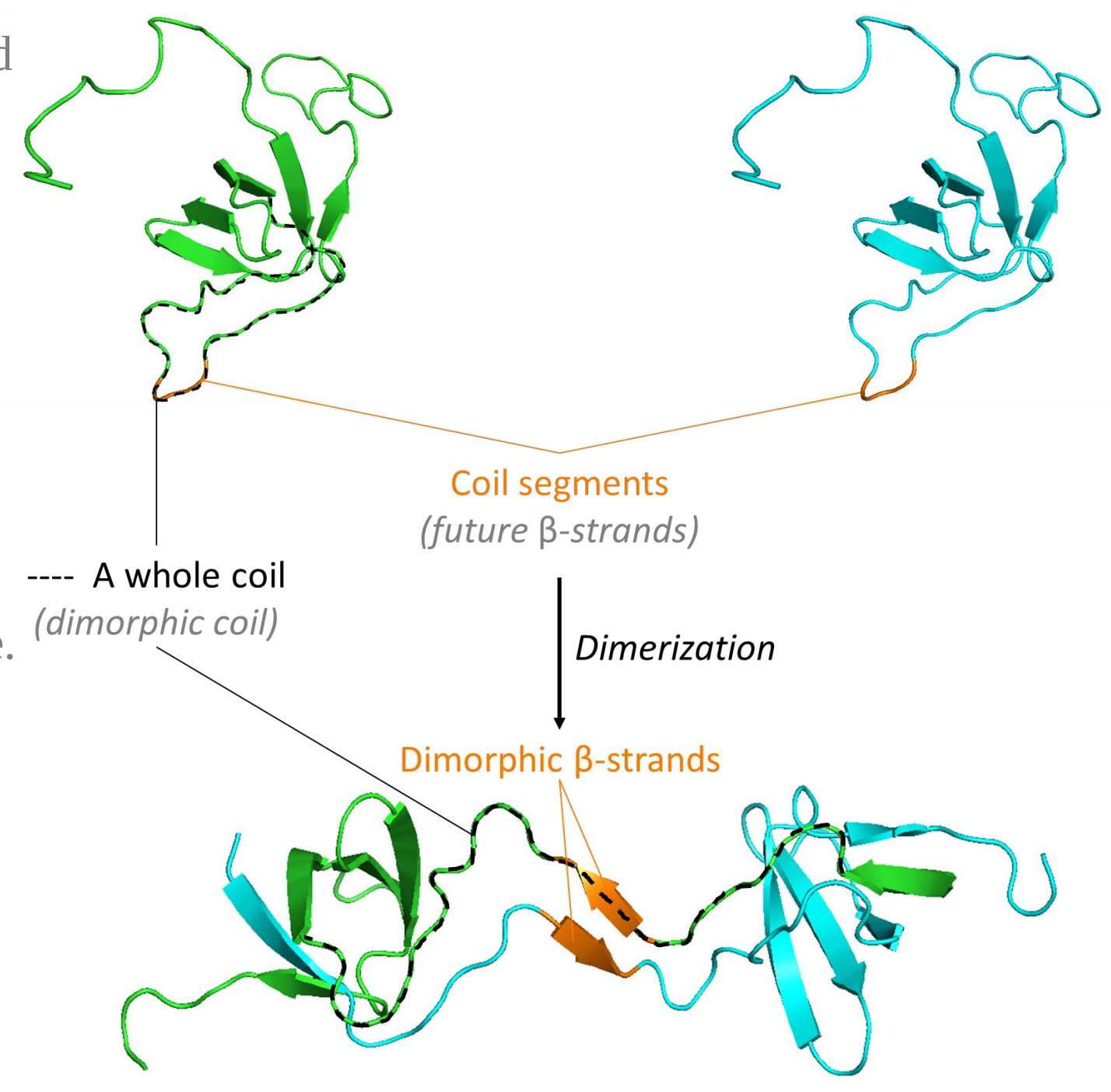
[1] Department of Bioengineering, Polytech Nice Engineering School, Côte d'Azur University, France.

[2] Department of Computer Science & Mathematics, Kingston University London, United Kingdom.

Contact: gabriel.flores-lipa@etu.univ-cotedazur.fr; j.nebel@kingston.ac.uk

Dimorphics and their significance

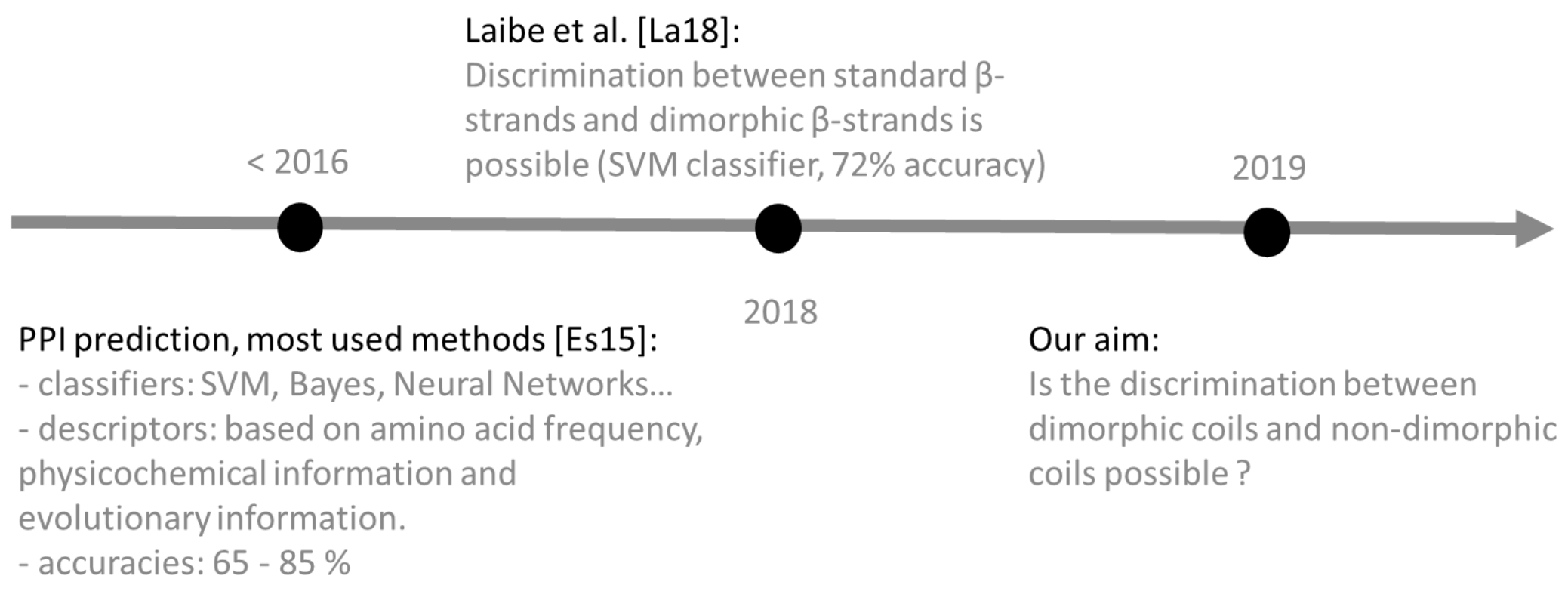
In 2018, *Laibe et al.* described a new class of protein segments – dimorphics - that adopt a coil conformation when the protein is in a monomeric form and a β-strand conformation after dimerization [La18]. Their specific characteristics suggest their *in-silico* identification may be possible. If it were the case, this would provide a simple approach to predict new Protein-Protein Interactions (PPIs), which may contribute to the development of new drugs, therapies and bioprocesses.



1

State of the art and objectives

PPI prediction can be considered as a classification machine learning problem. In this context, the following time-line shows common methodologies for *in-silico* PPI identification and progress towards dimorphic segment prediction.



Laibe et al. [La18]:
Discrimination between standard β-strands and dimorphic β-strands is possible (SVM classifier, 72% accuracy)

PPI prediction, most used methods [Es15]:
- classifiers: SVM, Bayes, Neural Networks…
- descriptors: based on amino acid frequency, physicochemical information and evolutionary information.
- accuracies: 65 - 85 %

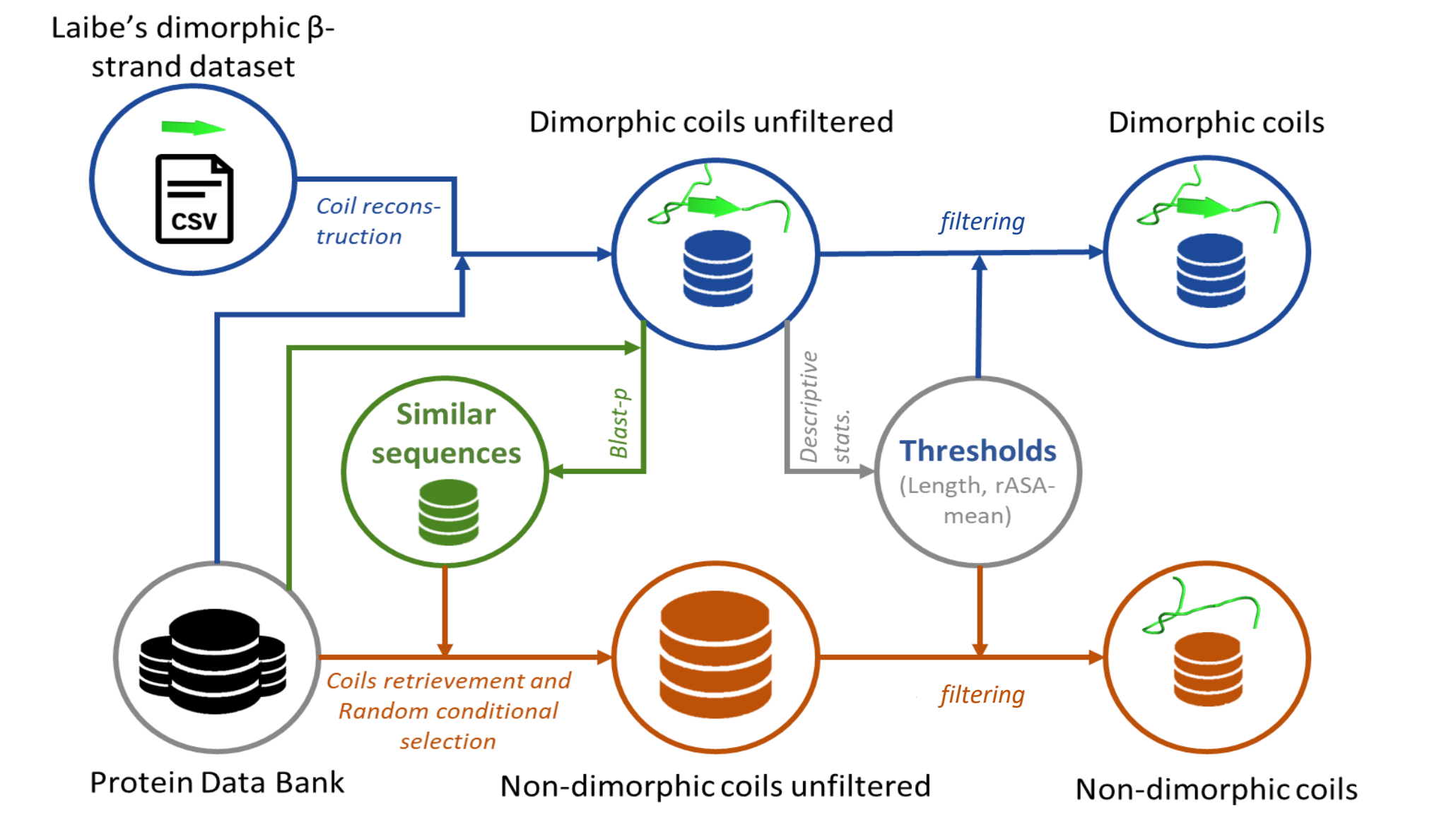
Our aim:
Is the discrimination between dimorphic coils and non-dimorphic coils possible ?

Objectives :

1. To construct a training dataset (dimorphic and non-dimorphic coils)
2. To construct a classifier (Machine Learning algorithm)
3. To analyze properties of proteins containing dimorphic-like sequences.

2

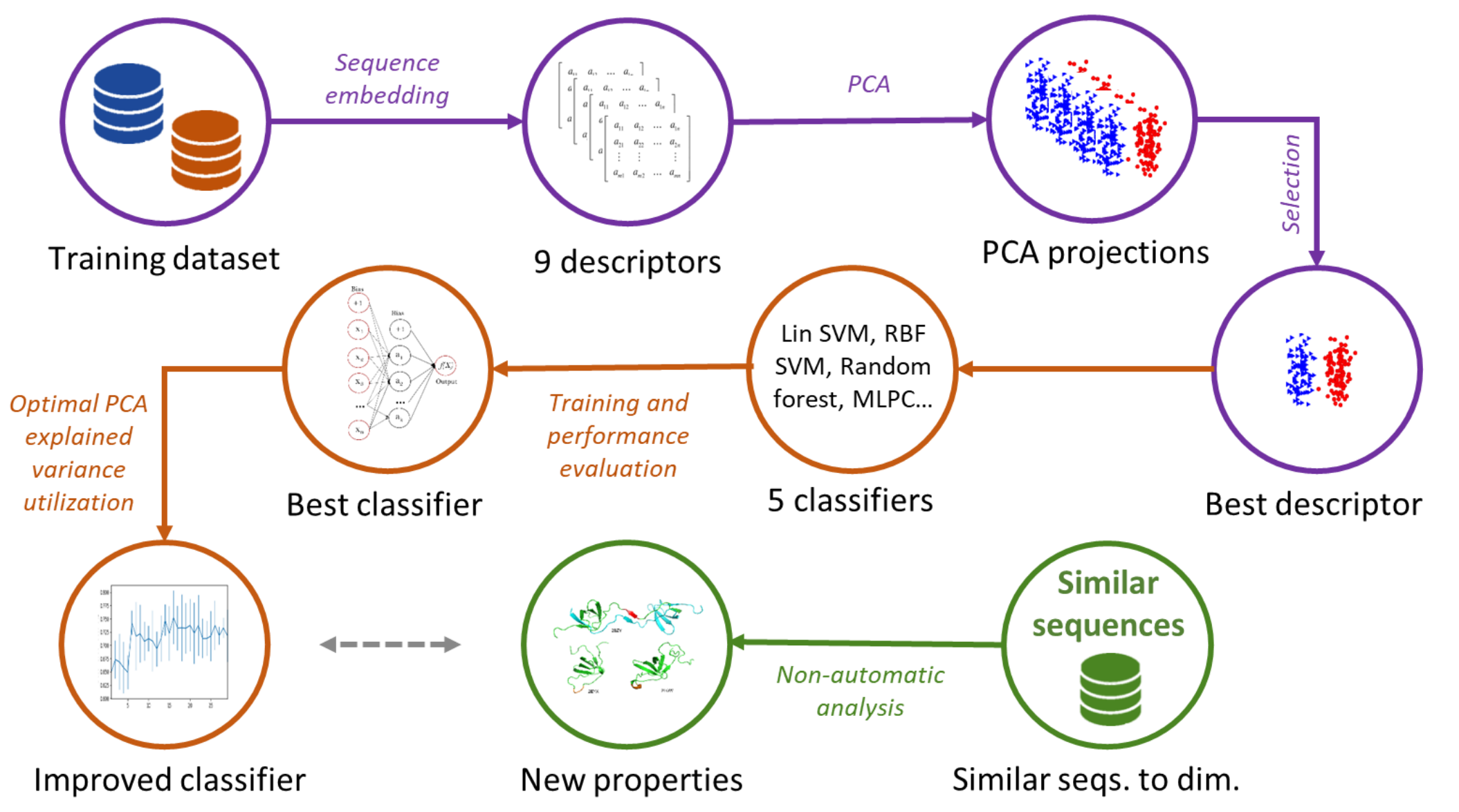
Training dataset construction


Laibe's dimorphic β-strand dataset
CSV
Coil recons-truction
Dimorphic coils unfiltered
filtering
Dimorphic coils
Blast-p
Similar sequences
Descriptive stats.
Thresholds (Length, rASA-mean)
Protein Data Bank
Coils retrievement and Random conditional selection
Non-dimorphic coils unfiltered
filtering
Non-dimorphic coils

The training dataset was constructed using the dimorphic dataset provided by *Laibe et al.* [La18] and the Protein Data Bank (PDB). Following extraction of dimorphic segments, dimorphic-like sequences were searched in the PDB using Blast-p. Such sequences were excluded from the non-dimorphic dataset. The dimorphic and non-dimorphic datasets were then filtered so that their distributions are consistent with the average length and mean relative Accessibility Surface Area (rASA-mean) of known dimorphics.

3

Classifier construction & non-automatic analysis

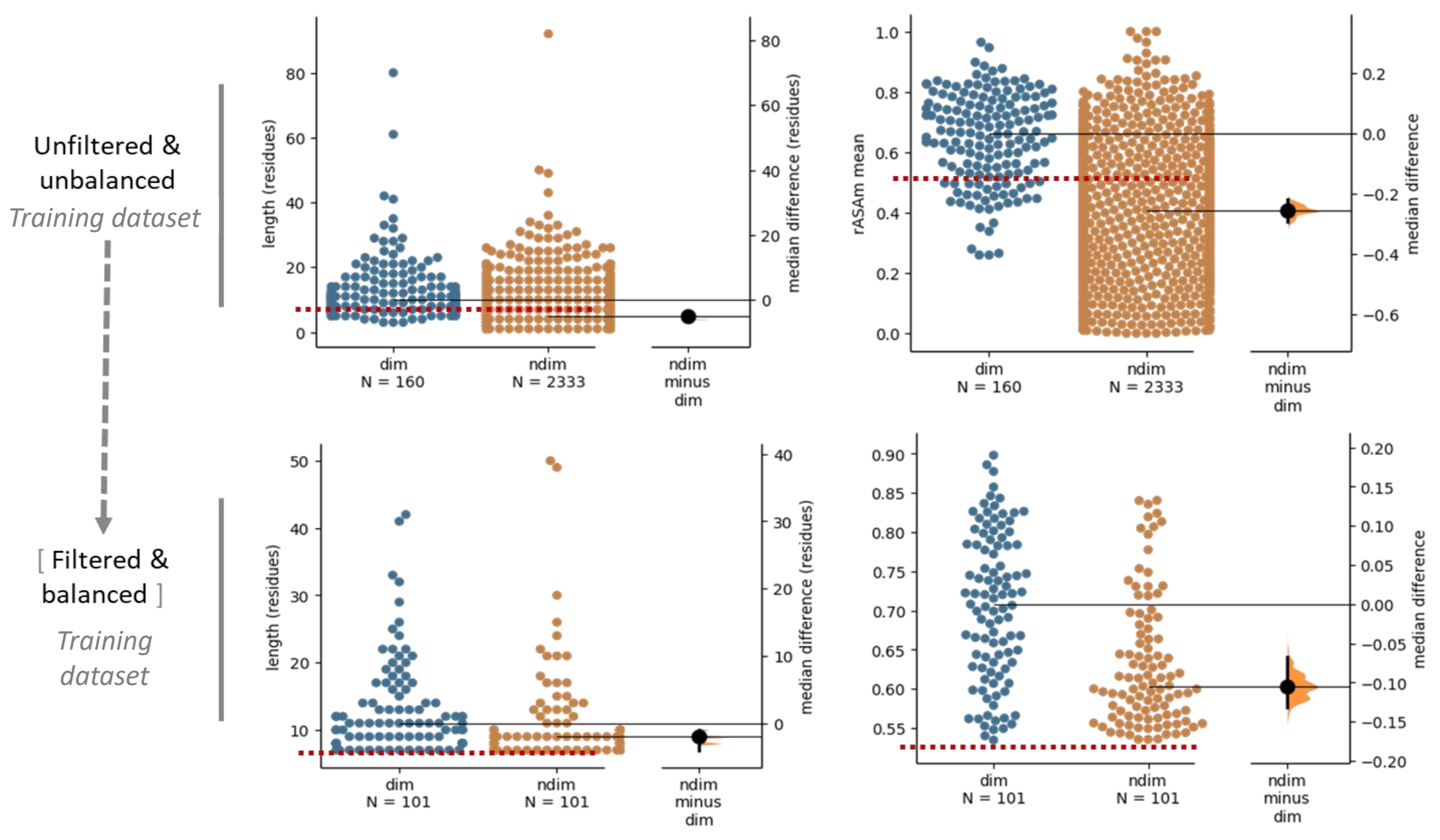


9 descriptors (vector representations of coil sequences) were computed using amino acids profile and physico-chemicals properties. The best descriptor was selected using Principal Component Analysis (PCA) projections. Using such descriptor, 5 different classifiers were trained. The best performing classifier was improved optimally using remaining PCA principal components of the best descriptor. Dimorphic-like sequences were analysed non-automatically. This helped to better understand final performance.

4

A homogeneous & balanced training dataset

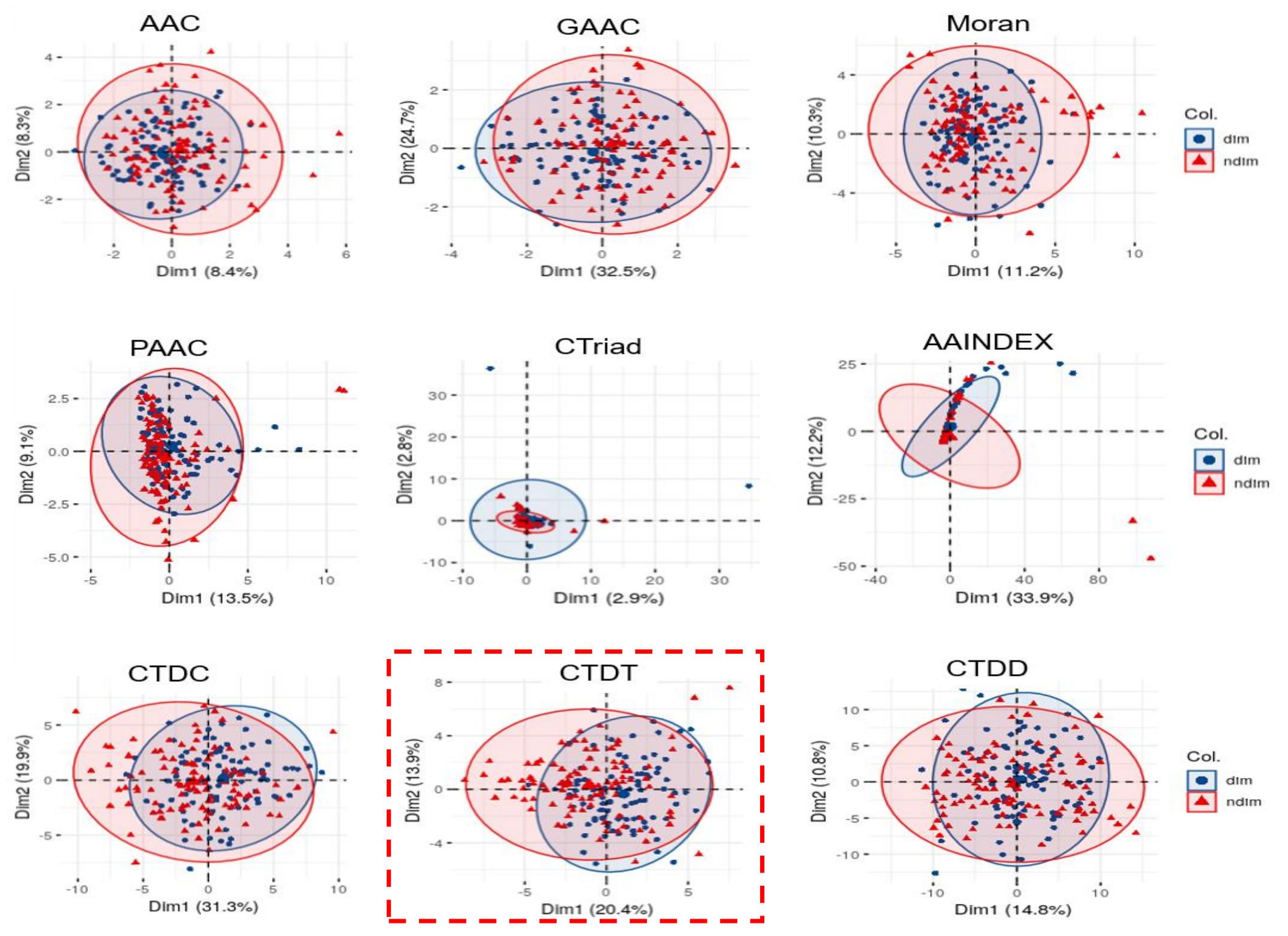
Using representative length and rASA-mean values (25 percentile in dimorphic segment distribution) as thresholds, both dimorphic and non-dimorphic datasets were filtered. The resulting training datasets comprise the same number of sequences, and similar length and rASA-mean distributions.



5

The best sequence descriptor

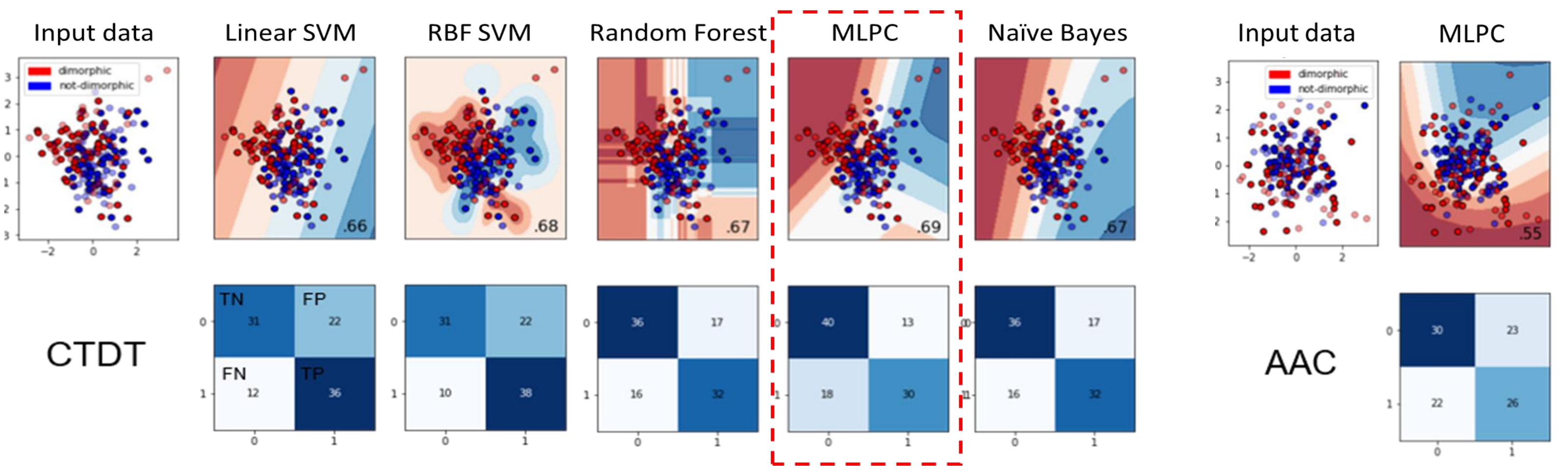
As illustrated below, among the 9 computed descriptors, the Composition Transition Distribution – Transition (CTDT) descriptor separates better dimorphic from non-dimorphic projections when using the 2 first Principal Components.



6

The best classifier

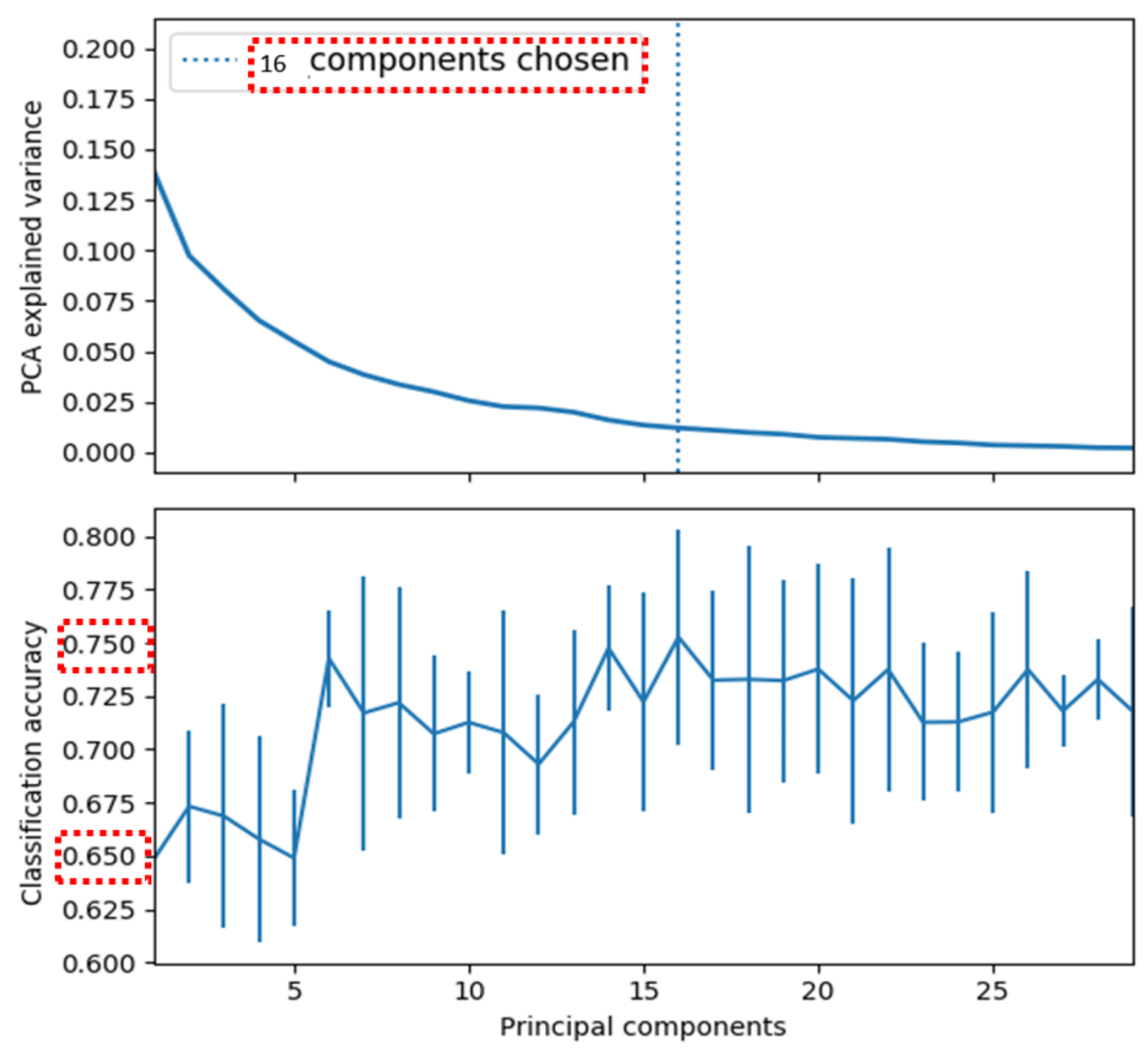
Using the 2 first PCA principal components of the CTDT descriptor, 5 classifiers were built. Among them, Multi-Layer Perceptron Classifier (MLPC) had the highest accuracy (69%), a balanced confusion matrix (high True Positive and True Negative rates, and low False positive and False Negative rates) and low overfitting, see results below. Performance of the Amino Acid Composition (AAC) descriptor is shown as control.



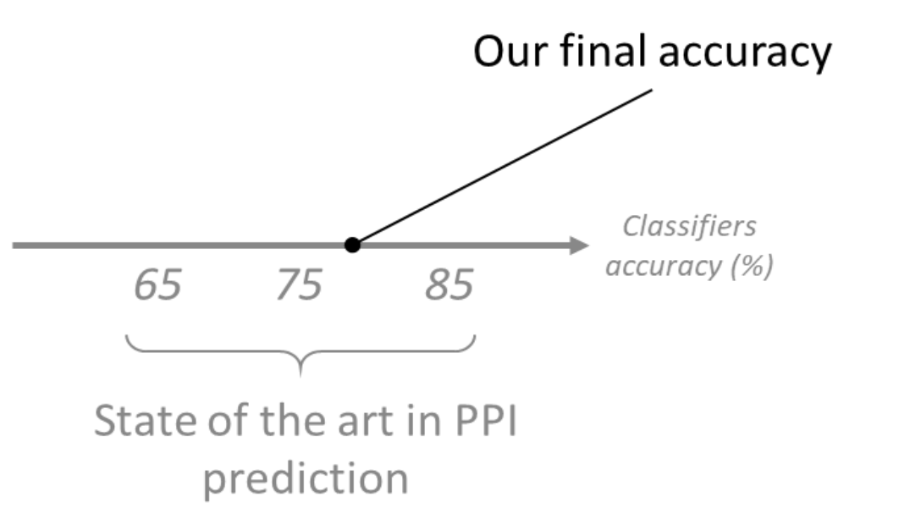
7

Final classifier improvement

Using the 2 first CTDT principal components and MLPC, the remaining explained variance from the others 14 CTDT principal components was optimally used to improve MLPC. This increased notably MLPC performance to 75% accuracy Consequently, usage of this classifier could contribute to identify new dimorphics in the PDB.



Mean +/- ds (N = 50)		Predicted class	
		non-dimorphic coil	dimorphic coil
Actual class	non-dimorphic coil	TN: 19.25 +/- 2.77	FN: 6.0 +/- 2.0
	dimorphic coils	FN: 6.0 +/- 1.87	TP: 19.25 +/- 2.77



8

Why only 75% accuracy? Clues from the non-automatic analysis…

Results from the manual analysis show that standard β-strands could also become dimorphic β-strands (case of the Proteinase Inhibitor from Nicotiana alata (NaProPI) synthesis), and that environmental conditions can dramatically affect a segment's secondary structure: in the case of the trypsin inhibitor C-terminal domain, **a trimorphic segment** was identified!.

NaProPI precursor is a repeated-domain protein. Each domain contains the A and B linked sub-domains. The A sub-domain contains b3 β-strand, and the B sub-domain contains b1 and b2 β-strands. After precursor circularization and proteolysis, domains separate. Monomers and a dimer are formed (c2). This dimer contains 2 β-strands (b3, from A sub-domain, and b1 from B sub-domain) that are dimorphic in the dimer (1QH2) and standard in monomers (e.g. 17IH).

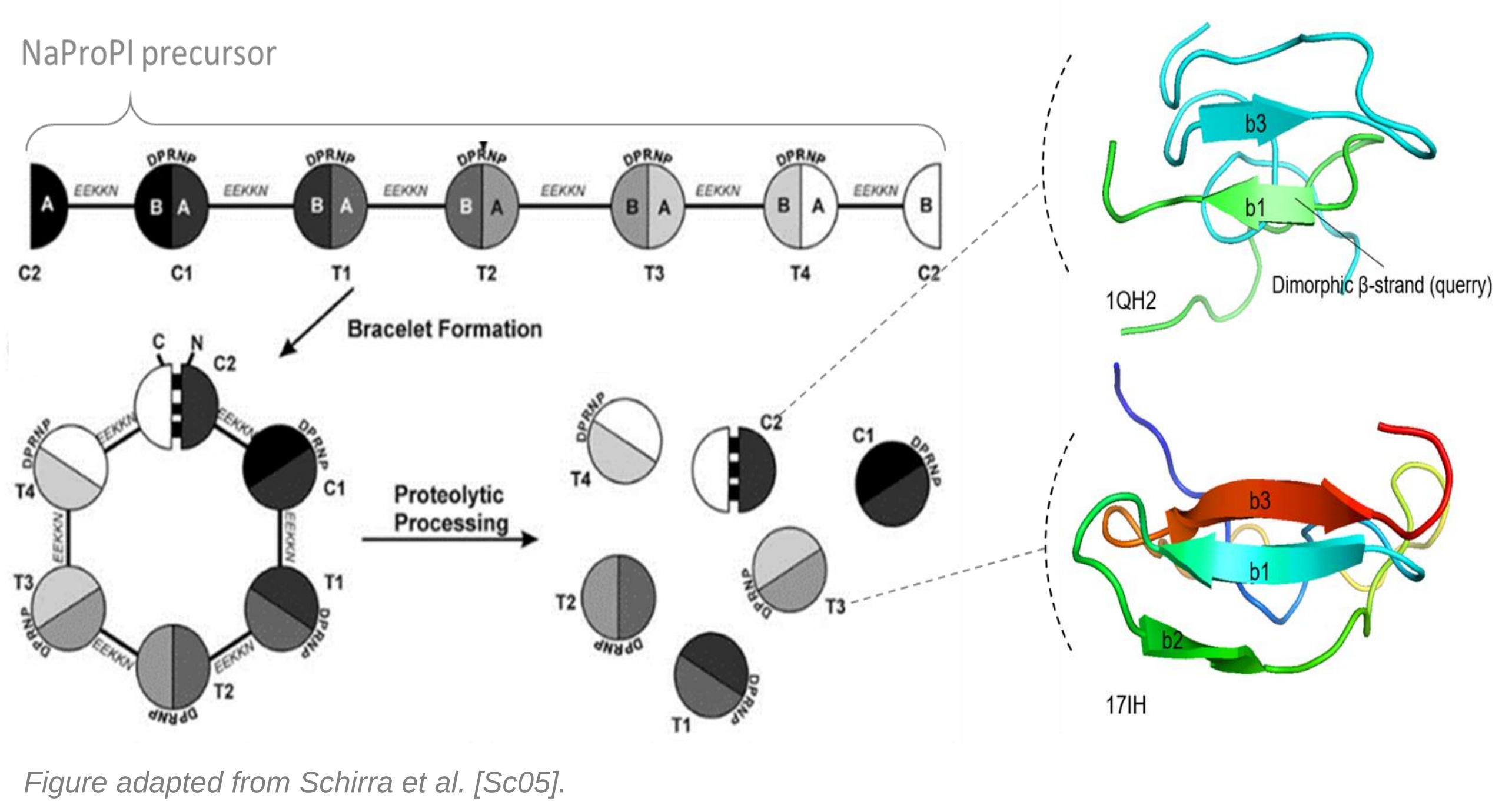
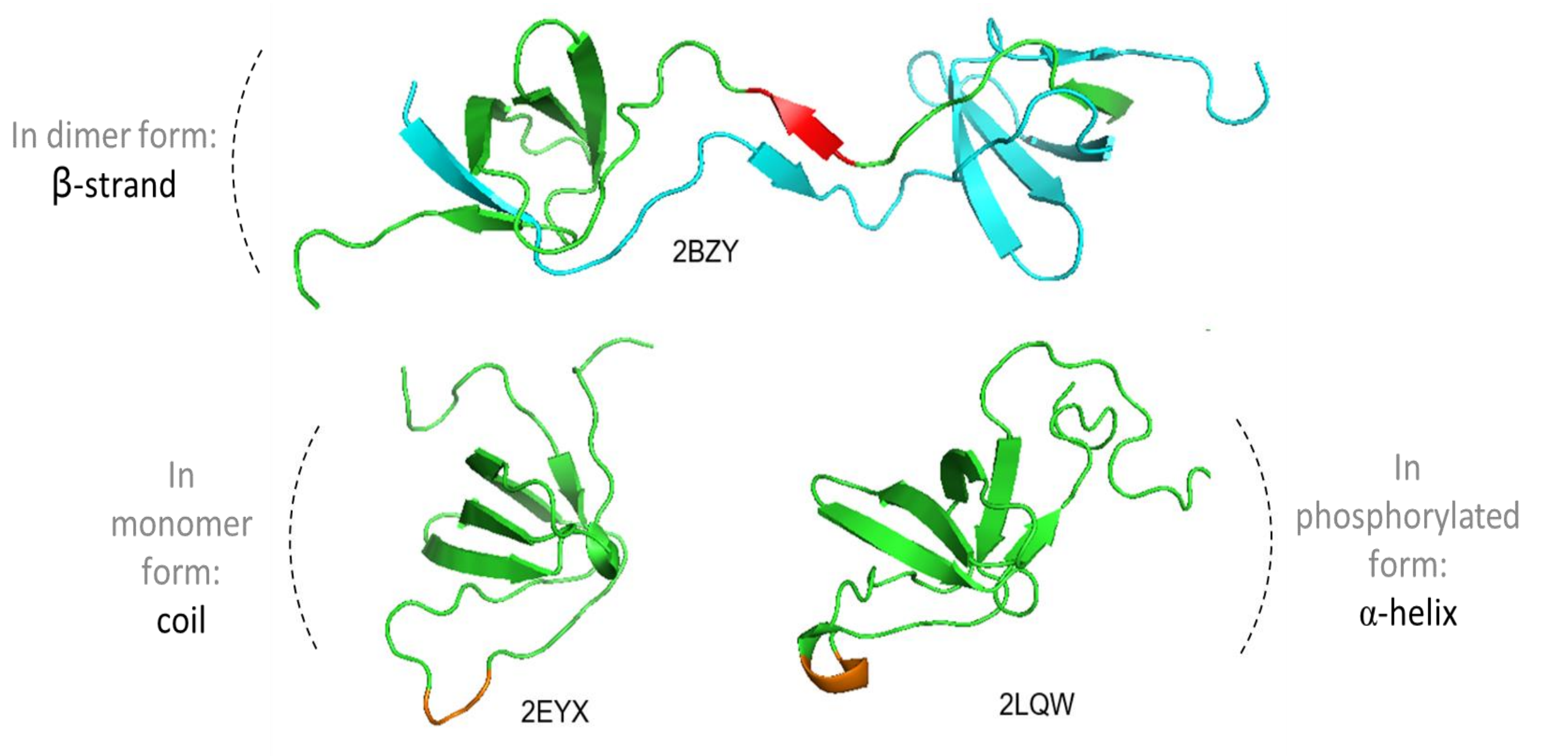


Figure adapted from Schirra et al. [Sc05].

C-terminal domain of a trypsin inhibitor (2EYX) can form homodimers involving 2 dimorphic β-strands (2BZY). Interestingly the same sequence (red in 2BZY and orange in 2EYX and 2LQW) can form a α-helix when trypsin inhibitor is phosphorylated. Thus, the same segment can adopt three stable conformations: coil, α-helix and β-strand.



Dimorphics are **fantastic**!

9

Conclusion and perspectives

Is the discrimination between dimorphic and non-dimorphic segments possible? Yes (75% accuracy using a MLPC classifier and 16 CTDT principal components). However, this study also shows that the primary structure, i.e. the sequence, does not contain all the required information for prediction. Indeed, both post-translational protein modifications and environmental conditions affect the secondary structure that such segments adopt.
How to improve predictions further? (i) gathering of more insight by analysis of misclassified coils, and (ii) refinement of chosen descriptor by adding extra information such as sequence polarity distribution and environmental conditions.

References

[La18] J. Laibe, A. Carey, M. Broutin, S. Guiglion, B. Pierscionek, and J.-C. Nebel. Coil conversion to β-strand induced by dimerization. Proteins: Structure, Function, and Bioinformatics, 86(12):1221-1230, 2018.

[Es15] R. Esmaielbeiki, K. Krawczyk, B. Knapp, J.-C. Nebel, and C. M. Deane. Progress and challenges in predicting protein interfaces. Briefings in bioinformatics, 17(1):117131, 2015.

[Sc05] H. J. Schirra and D. J. Craik. Structure and folding of potato type ii proteinase inhibitors: circular permutation and intramolecular domain swapping. Protein and peptide letters, 12(5):421431, 2005.

Box code colour: blue: introduction, orange: methods, green: results and discussion, red: conclusion