Facilitating the study of peptides sequences using machine learning strategies and bioinformatics tools.

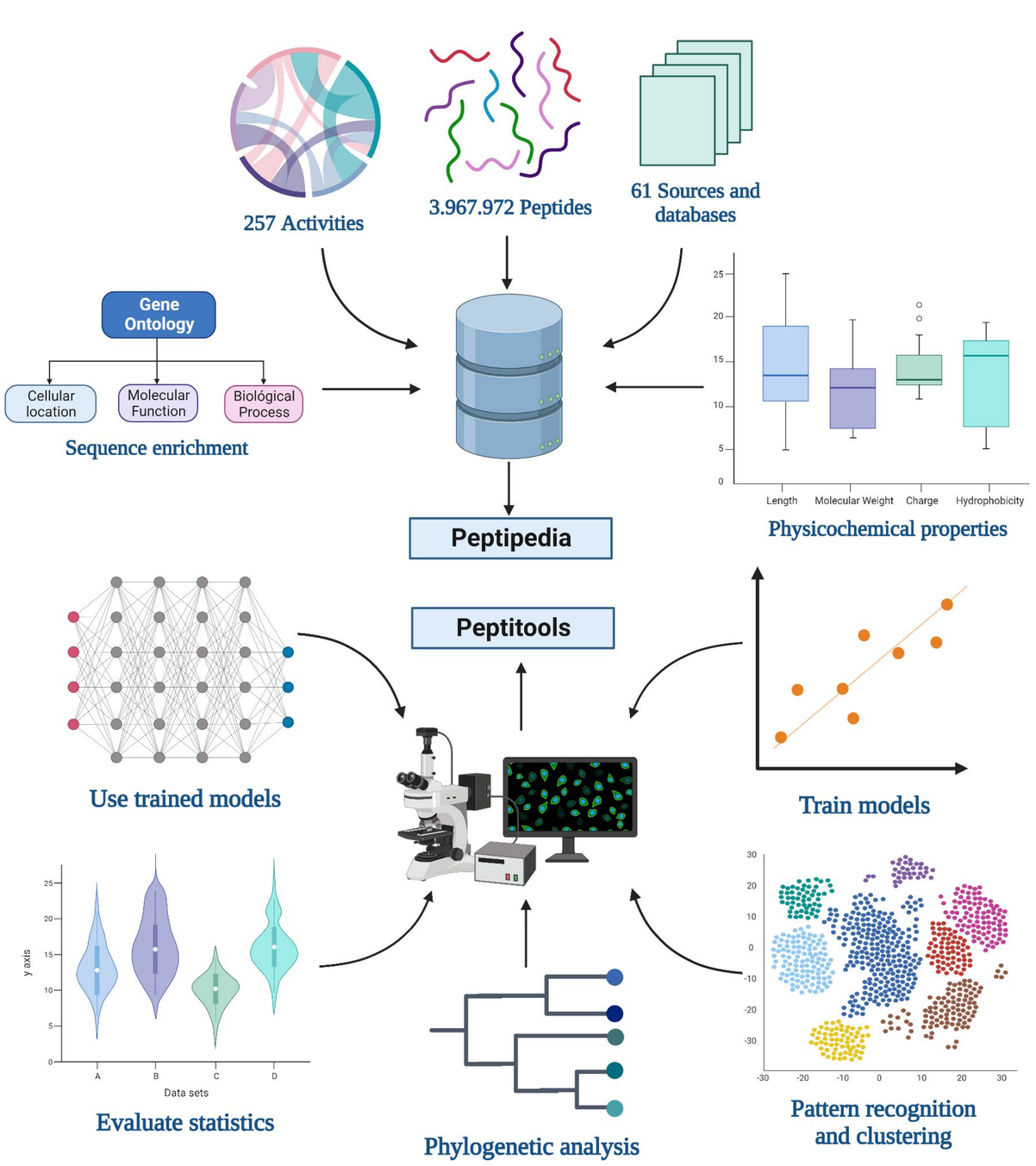
Gabriel Cabas Mora[1], Álvaro Olivera-Nappa[2] , David Medina Ortiz[1,2]
[1]Departamento de Ingeniería en Computación, Universidad de Magallanes, [2]Centre for Biotechnology and Bioengineering, Universidad de Chile
Contact: david.medina@umag.cl or gabriel.cabas@umag.cl

Motivation and objectives

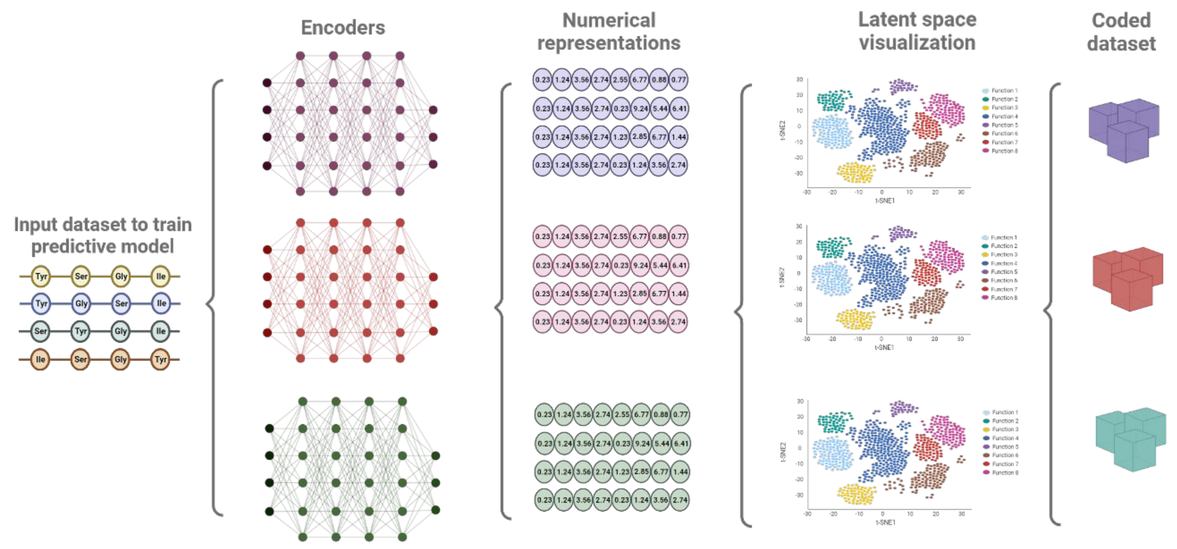
Peptides are relevant in several biotechnology applications. These molecules have different biological activities, or desired properties. In particular, the peptides are attractive as therapeutic agents. New research has fostered the exponential increase of these molecules in common or specific databases. However, there needs to be more user-friendly tools to make up for the lack of bioinformatics or machine learning skills to study peptide sequences, and the disorder and scatter of information.

Peptipedia database has the most significant number of peptides with reported biological activity

In this work, we significantly updated our peptide sequence database, Peptipedia. We incorporated 61 data sources, 257 activities, and more than 95,000 peptide sequences with reported biological activity. In addition, enrichment analysis via Gene Ontology (GO) and Pfam functional domains were incorporated to complement the information on the functional and physicochemical characteristics enabled for each peptide. Besides, we updated the services of our platform, incorporating statistical evaluation methods, functional predictions of GO, and functional domains via Pfam. Furthermore, large language models have been enabled to apply numerical representation strategies, which, combined with pattern identification methods and predictive model development, facilitate the application of machine learning techniques for studying peptide sequences.

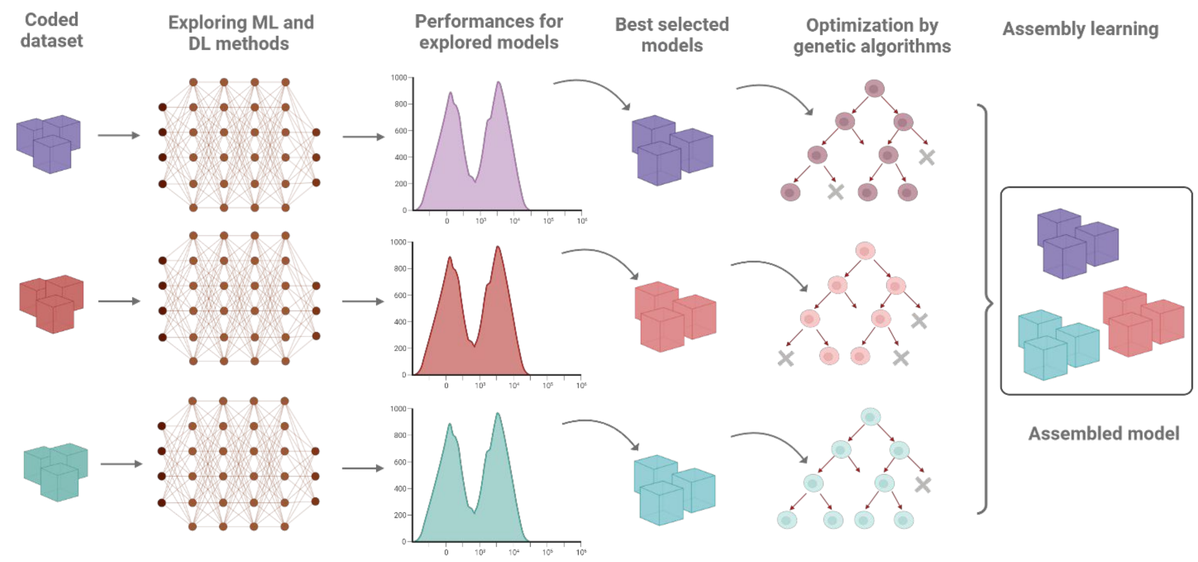


Using a generalizable sequence-based approach to implement relevant predictive models for peptide sequences



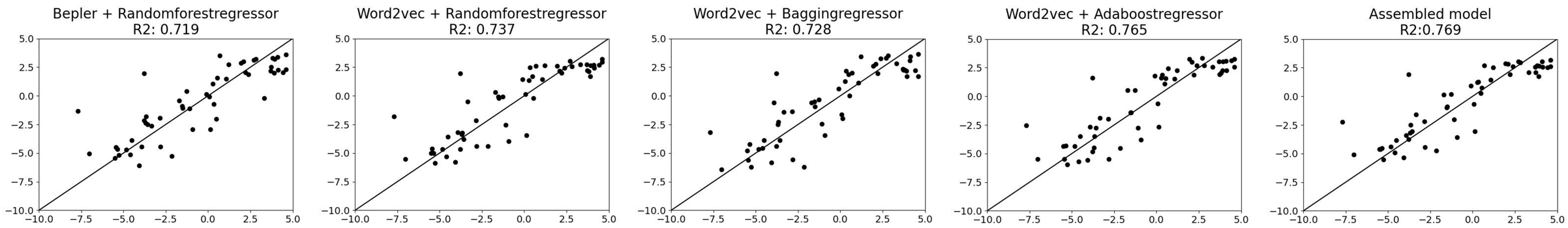
Different numerical representation strategies were explored for protein sequences. Amino acid encoding techniques were applied using physicochemical properties extracted from the AAIndex database. Besides, the Fast Fourier transform was employed to represent the amino acid coding results as frequency signals. In both cases, zero-padding strategies were utilized to ensure equal vector size. Finally, protein language models were applied to represent numerically the enzyme sequences. This work used different pre-trained models available in the literature, such as Bert, bepler, ESM1b, and Prottrans.

We combine Large Language Models and Assembled learning to develop accurate predictive models



Machine learning algorithms and deep learning architectures are explored using default hyperparameters and evaluating via classic performances. Then, statistical methods are used to select the best combinations of algorithms and numerical representations strategies, whose hyperparameters are optimized via genetic algorithms. Finally, assembled learning strategies are applied to build a single predictive system.

Using the assembled models it was implemented a predictive model for the IC50 activity in Antiviral peptides, achieving a 0.77 or R^2 coefficient



Funding

The authors acknowledge funding by the MAG-2095 project, Ministry of Education, Chile. DMO acknowledges ANID for the project "SUBVENCIÓN A LA INSTALACIÓN EN LA ACADEMIA CONVOCATORIA AÑO 2022", Folio 85220004. DMO and AON gratefully acknowledge support from the Centre for Biotechnology and Bioengineering - CeBiB (PIA project FB0001, Conicyt, Chile).