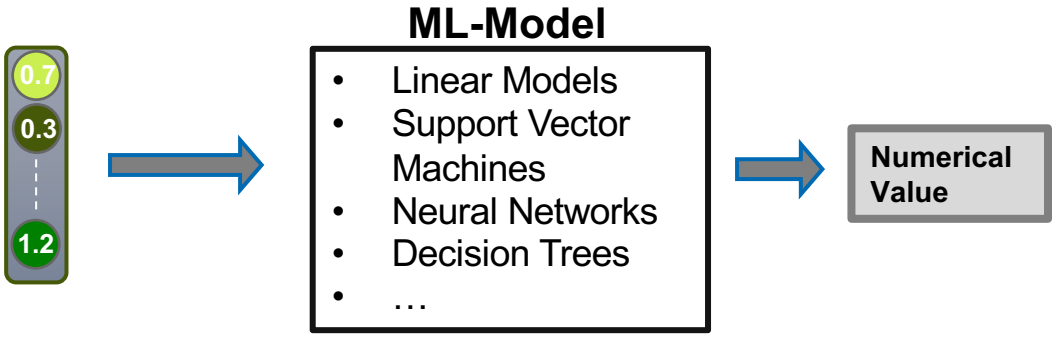
Heinrich Heine
Universität Düsseldorf

Traditional numerical representations for molecules

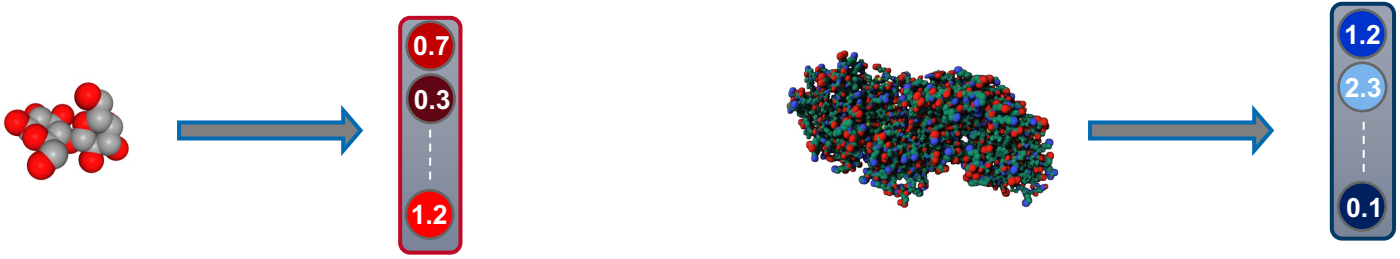
How can we make predictions for molecules?

Heinrich Heine Universität Düsseldorf

- Machine Learning models require numerical representations as input:

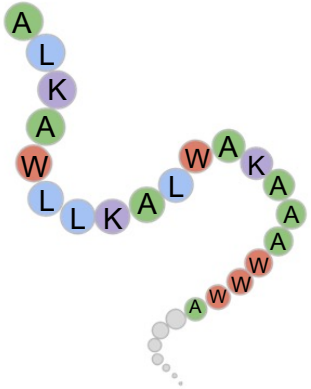


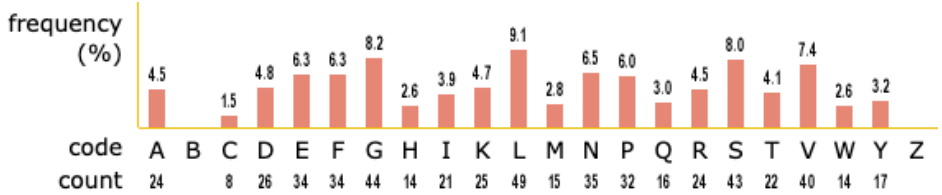
- We need numerical representations of molecules



Amino Acid Composition

- Amino Acid Composition (AAC):
  - Percentages of each amino acid in the sequence of that protein
  - Results in a vector $(f_1, f_2, \dots, f_{20})$
  - Example:





Number of residues = 537 Molecular weight = 60673.44

$$\vec{f} = (4.5, 0, 1.5, \dots, 3.2, 0)$$

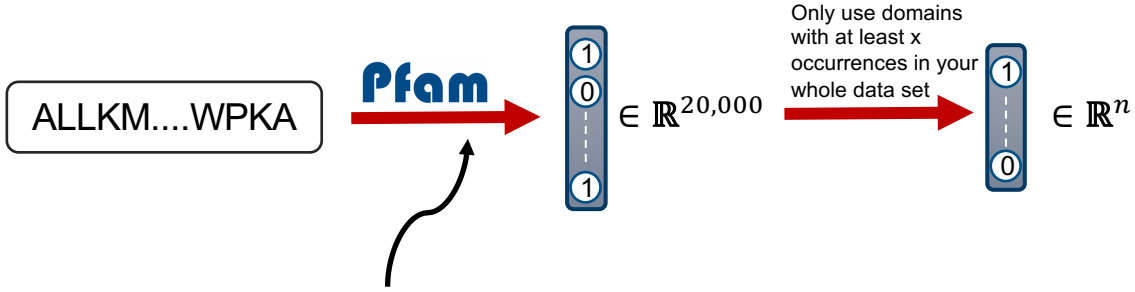
Bar plot copied from: https://proteopedia.org/wiki/index.php/Amino_acid_composition

k-mer Features

- k-mer (=n-gram) features:
  - Creating a list of all possible amino acid sequences of length k
  - Calculate the frequencies of of these strings in the protein amino acid sequence
  - Use the resulting features in a vector $(f_1, f_2, \dots, f_n),\ n = 20^k$
    - If n becomes too large, only use the most common substrings
  - For k = 1, we get the amino acid composition
  - Example for k = 3:
    - Calculating the frequency of each of the subsequences AAA, AAR, AAN, AAD, AAE, AAG, AAC, AAQ, AAH, …. (8000 possible 3-mer) within the protein amino acid sequence: $(f_1, f_2, \dots, f_{8000})$
    - These subsequences can be overlapping:
      - “AAAM” contains both AAA and AAM

Protein Domains

- Protein Domains:
  - Subsequences of proteins with specific structural or functional characteristics
  - Pfam database: 20k different protein domains



- The amino acid sequence of every protein can be mapped via the webservice of Pfam to functional domains

Pfam entries

Heinrich Heine Universität Düsseldorf

Summary

This is the summary of UniProt entry VAV_HUMAN (P15498).

Description:	Proto-oncogene vav
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606)
Length:	845 amino acids
Reference Proteome:	✓

Please note: when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

Pfam domains

This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains. **More...**

Download the data used to generate the domain graphic in JSON format.

Source	Domain	Start	End
Pfam	CH	1	121
low_complexity	n/a	41	50
disorder	n/a	128	129
disorder	n/a	140	141
disorder	n/a	160	161
disorder	n/a	173	177
disorder	n/a	179	180
Pfam	RhoGEF	198	371
low_complexity	n/a	355	366
Pfam	PH	403	504
Pfam	C1_1	516	568
disorder	n/a	568	588
Pfam	SH3_1	615	652
disorder	n/a	635	636
Pfam	SH2	671	745
Pfam	SH3_1	788	834

Show or hide domain scores.

Calponin homology domain

Calponin homology domain (or CH domain) is a family of actin binding domains found in both cytoskeletal proteins and signal transduction proteins.[2] The domain is about 100 amino acids in length and is composed of four alpha helices.[3] It comprises the following groups of actin-binding domains:

- Actinin-type (including spectrin, fimbrin, ABP-280)
- Calponin-type

A comprehensive review of proteins containing this type of actin-binding domains is given in.[4]

The CH domain is involved in actin binding in some members of the family. However, in calponins there is evidence that the CH domain is not involved in its actin binding activity.[5] Most proteins have two copies of the CH domain, however some proteins such as calponin and the human vav proto-oncogene (P15498) have only a single copy. The structure of an example CH domain has been determined using X-ray crystallography.[6]

Calponin homology (CH) domain