Pushing the limits in automated NMR structure determination

Tomáš Brázdil

In collaboration with Thomas Evangelidis, Jiří Filipovič, Jana Hozzová, Jaroslav Oľha, David Porteš, Kostas Tripsianes

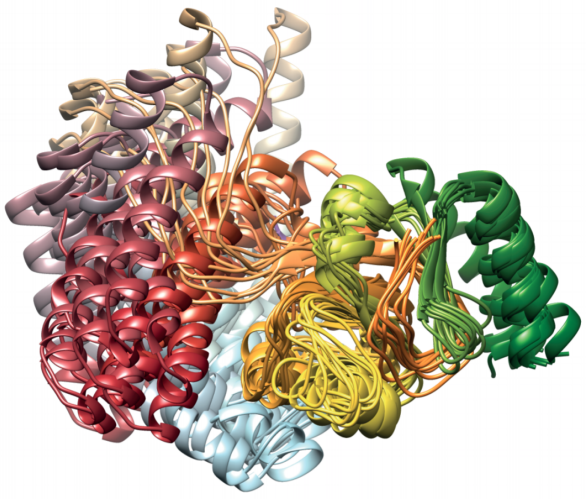
Masaryk University
Brno, Czech Republic

Proteins

- Biomolecules, long chains of amino acid residues.

  Size ranges from ~20 to ~30,000 amino acid residues.

- Perform a vast array of functions within organisms.
- Determining structure is important for understanding function.

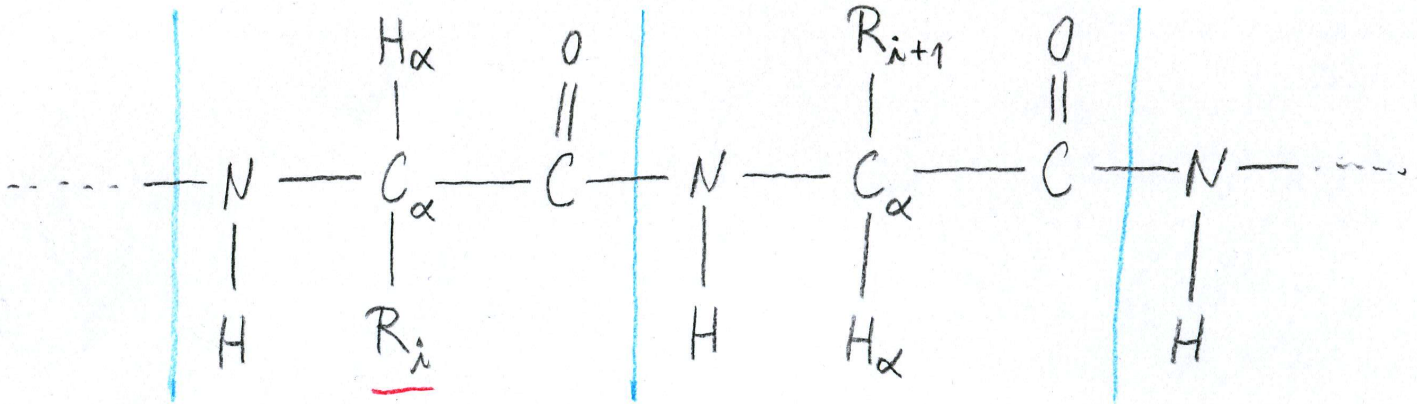


Several methods for protein structure determination: X-ray crystalography, **NMR spectroscopy**, Electron microscopy, ...

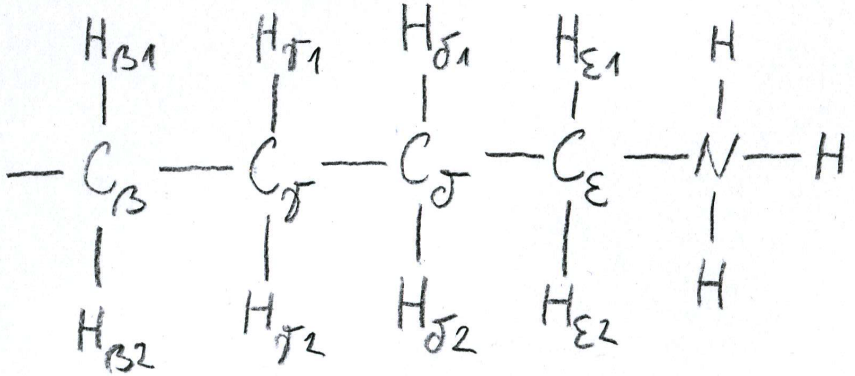
Advanced computational methods used to determine the shapes of the proteins.

Proteins

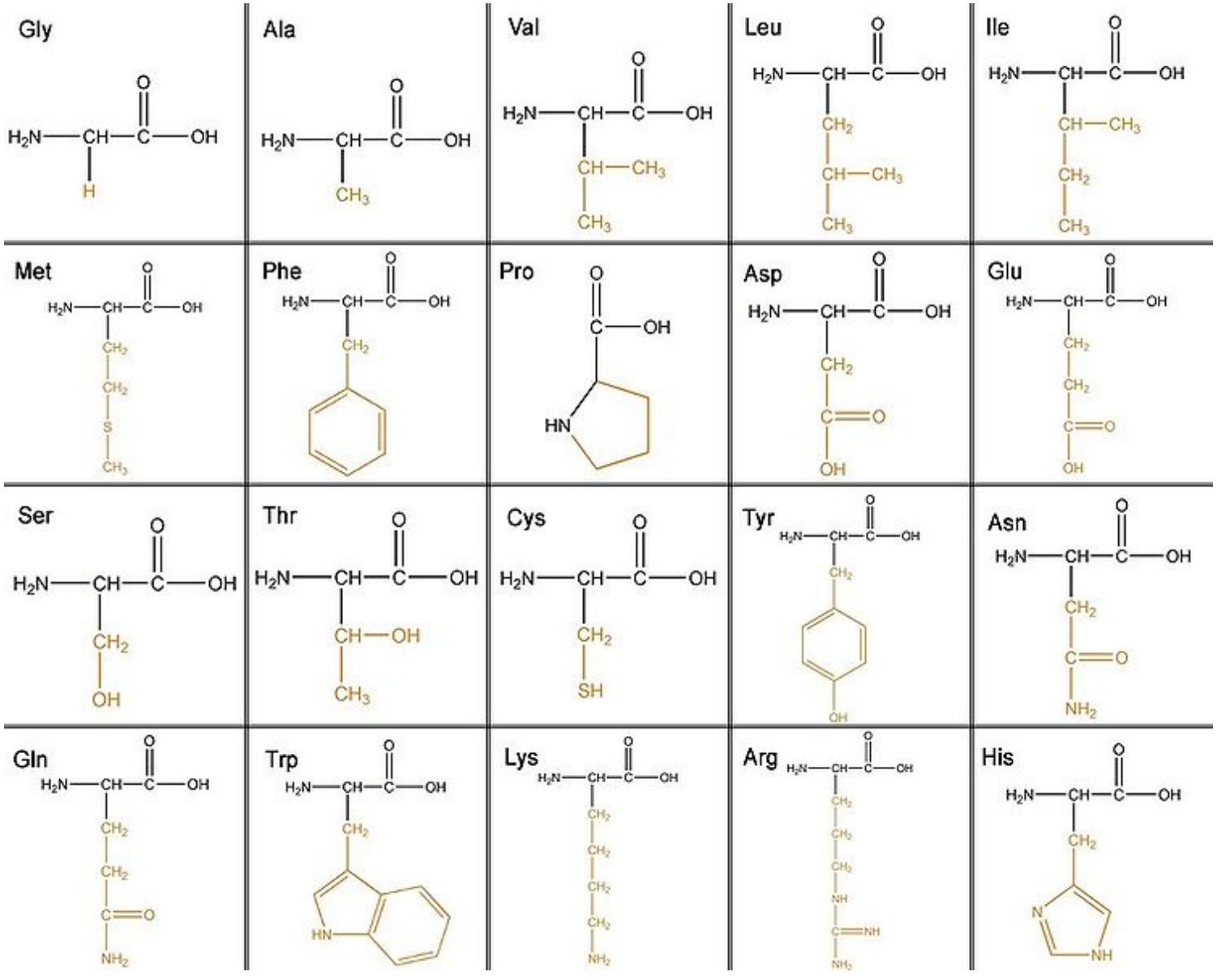
Sequences of amino acid residues.



R_i : LYS



Amino acids


Gly
Ala
Val
Leu
Ile
Met
Phe
Pro
Asp
Glu
Ser
Thr
Cys
Tyr
Asn
Gln
Trp
Lys
Arg
His

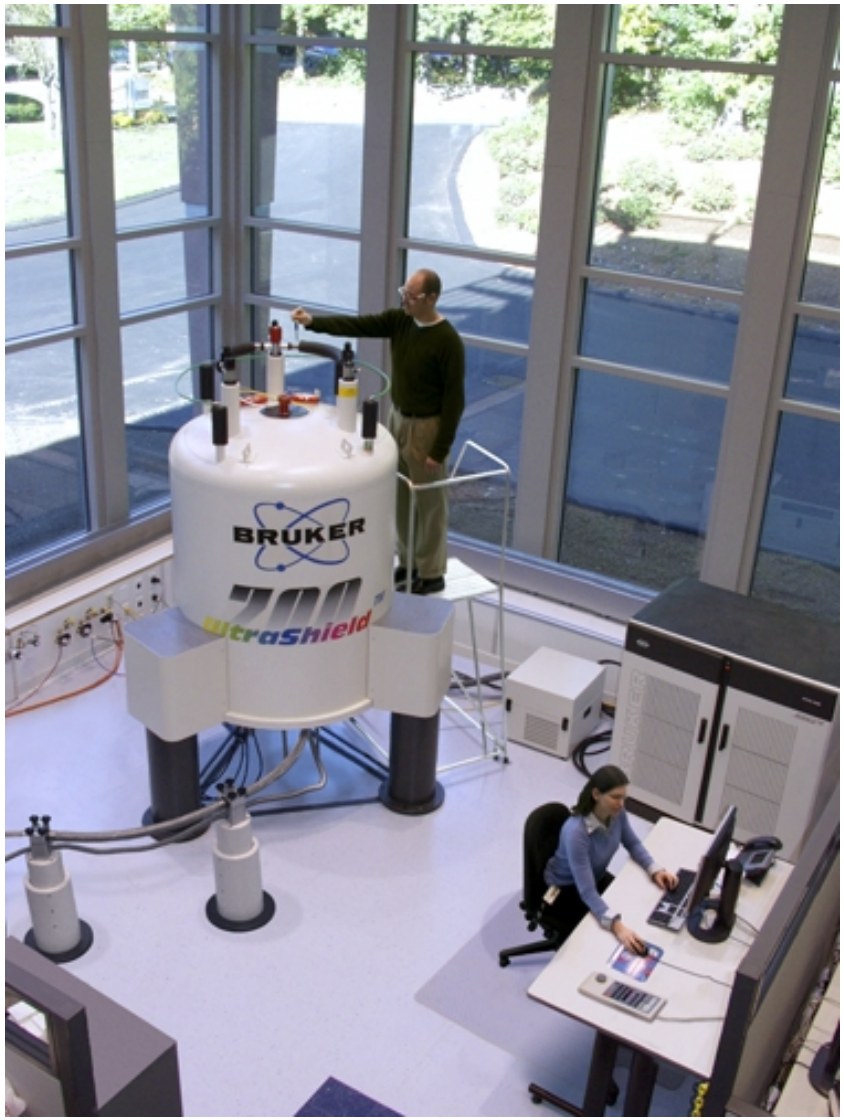
NMR spectroscopy

A method for determining structural and dynamic properties of molecules.

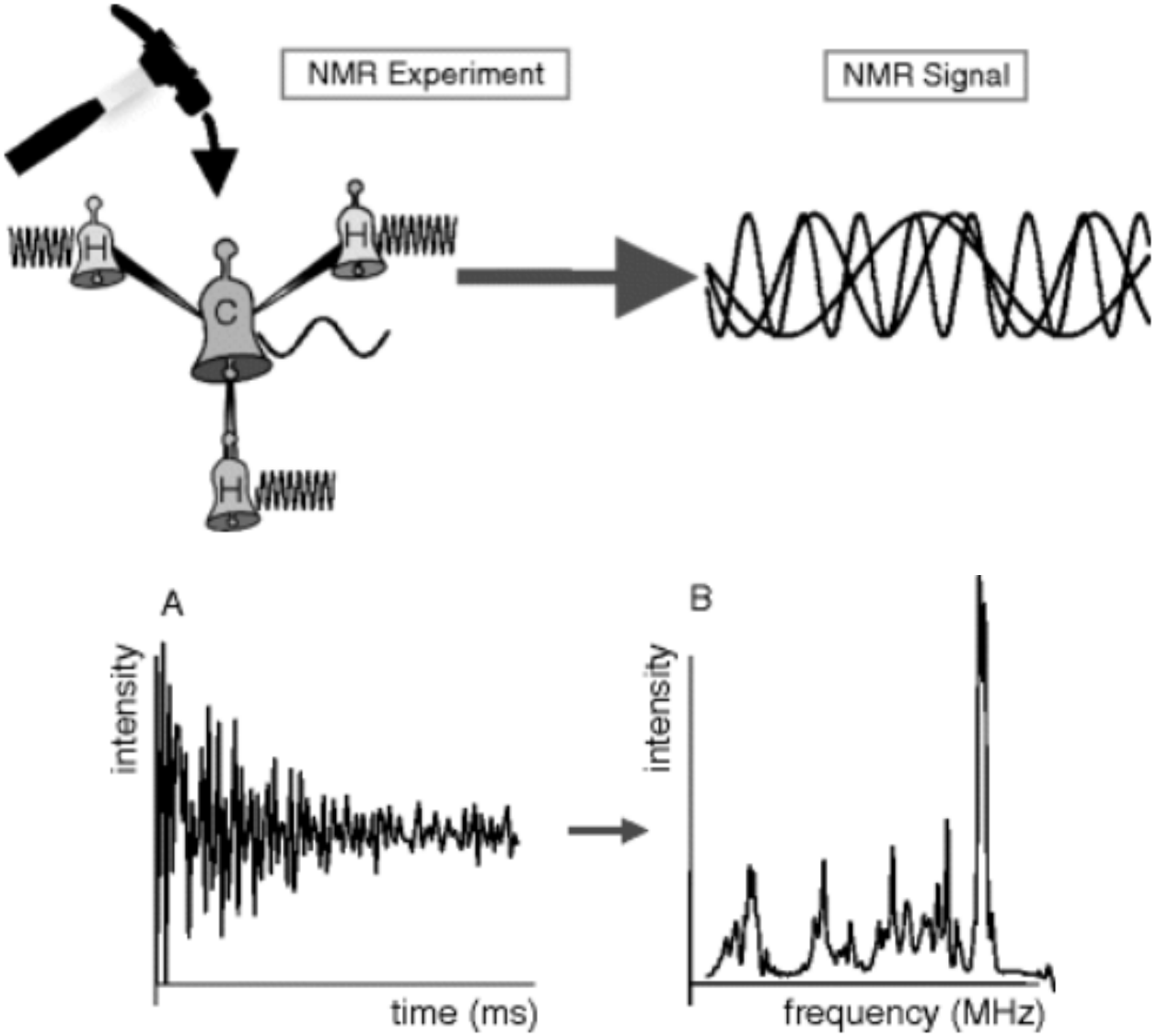
- atom distances
- bond angles
- rates of movement
- ...

Size zone: 10 to 40 kDa

- Cryo-em: >50 kDa only
- X-ray: Crystals needed
- NMR: Currently slow

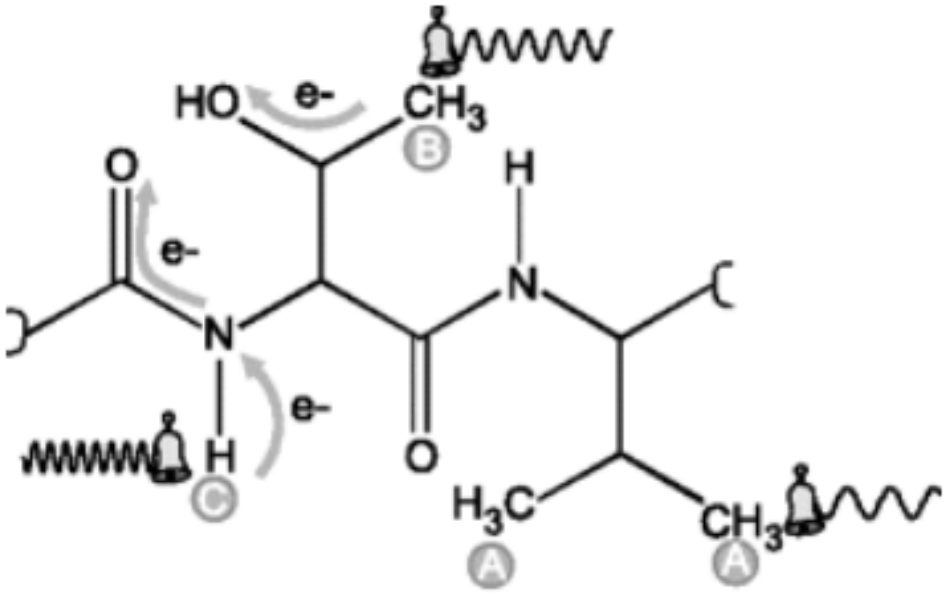


NMR experiment



Source of images: Pocket Guide to Biomolecular NMR. M. Doucleff, Springer 2011.

NMR experiment

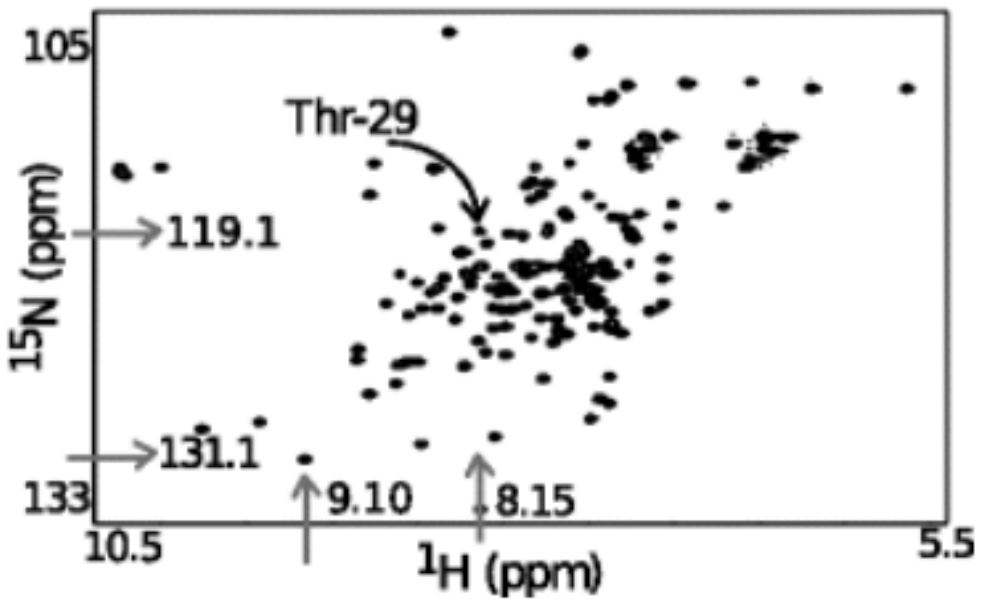


- The "ringing" frequency depends on the environment of the nucleus.
- chemical shift = relative "ringing" frequency of a nucleus (in ppm) w.r.t. the frequency of nuclei in TMS (tetramethylsilane)

Source of images: Pocket Guide to Biomolecular NMR. M. Doucleff, Springer 2011.

2D NMR spectroscopy

2D NMR spectroscopy



Source of images: Pocket Guide to Biomolecular NMR. M. Doucleff, Springer 2011.

4D HCNH NOESY

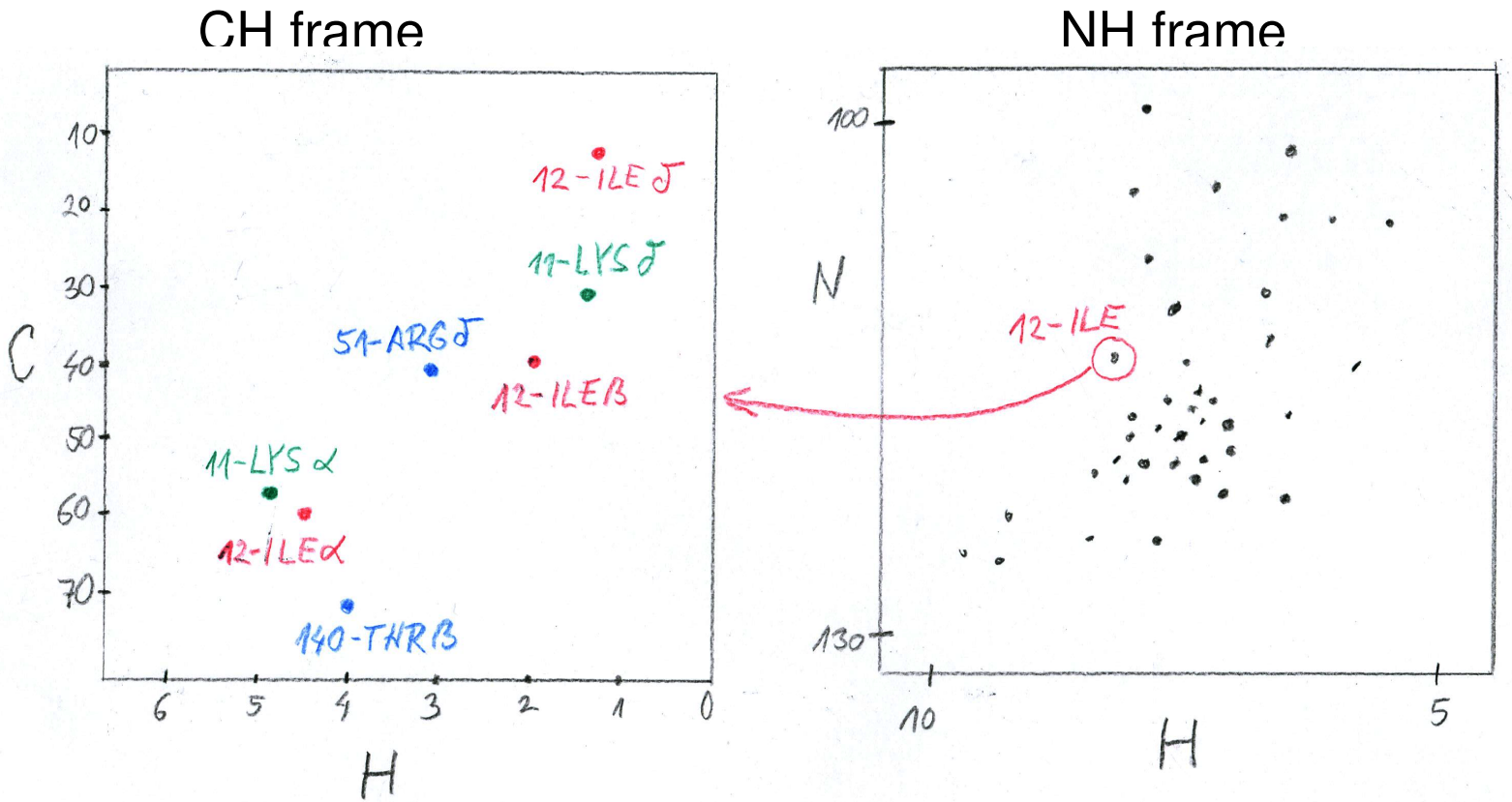
Consider a sequence of amino acid residues

4D HCNH NOESY

Consider a sequence of amino acid residues

.... 10-GLY 11-LYS 12-ILE 13-ARG 14-ALA....

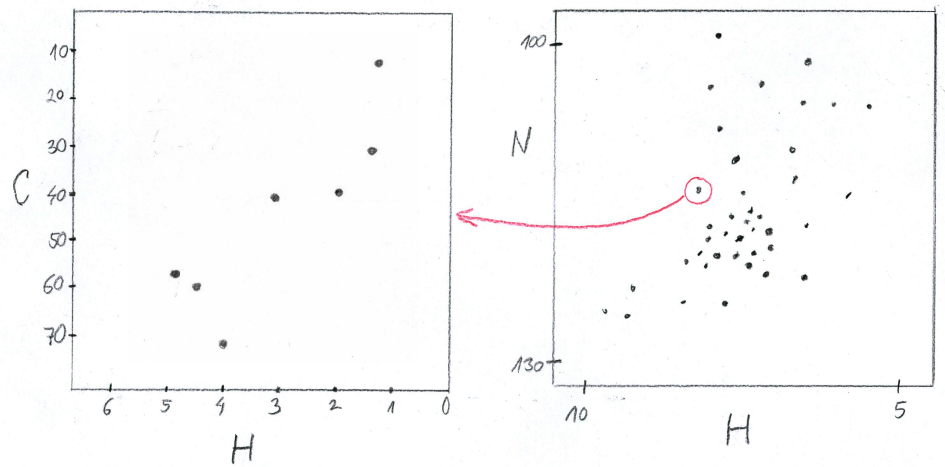
The 4D HCNH NOESY data look like this:



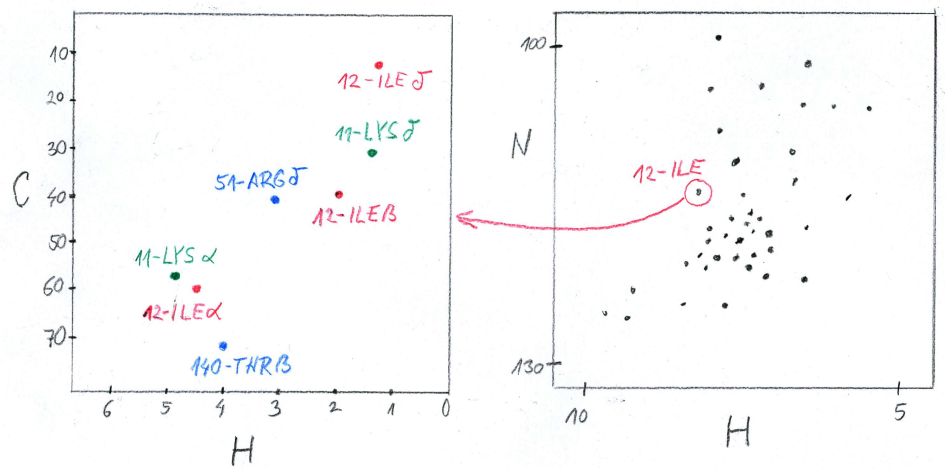
Each CH frame corresponds to one residue in the sequence.

The complete assignment problem

Given a protein sequence of amino acids and this



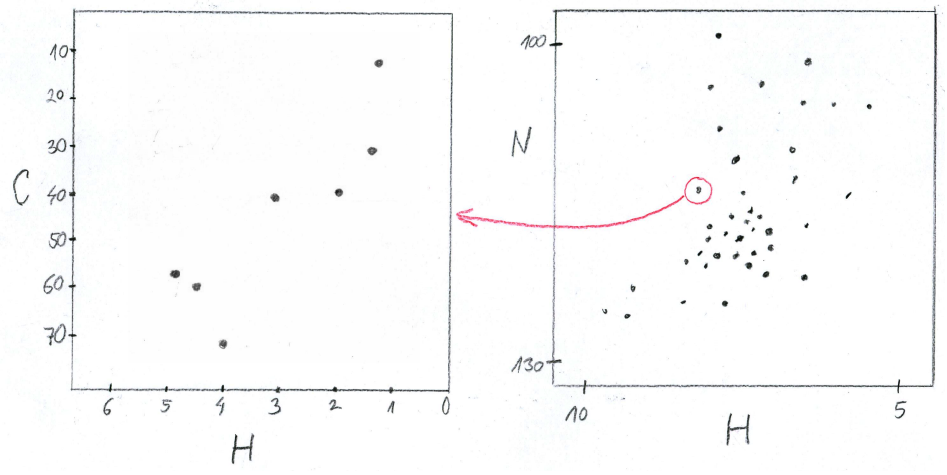
produce this



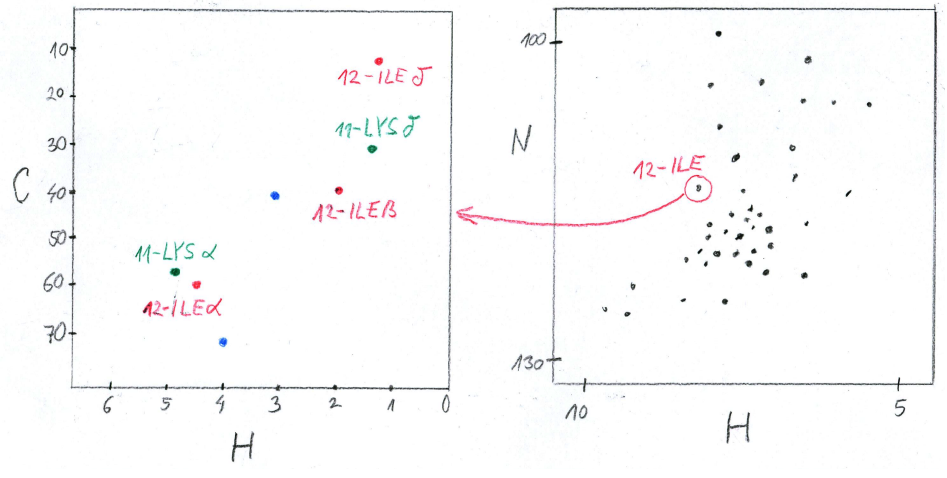
... and do it for all peaks in all NH and CH frames.

Our assignment problem

Given a protein sequence of amino acids and this



produce this



... and do it for all peaks in the NH frame.

The assignment problem solutions

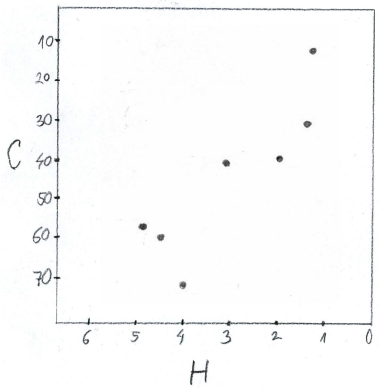
Standard solution of the assignment:

- Additional 6 – 10 3D experiments that are of no use for the structure determination.
- Assignment difficulty related to the size:
  - easy for <15 kDa
  - difficult for 15 - 20 kDa
  - heoric for >20 kDa
- Previous work:
  Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. Evangelidis et al. nature communications, 2018
  - 4D-CHAINS tool
  - Uses two spectra: 4D HCNH NOESY & HCNH TOCSY

Our goal: Compute the assignment based just on the single 4D HCNH NOESY spectrum.

The plan

Given a CH frame



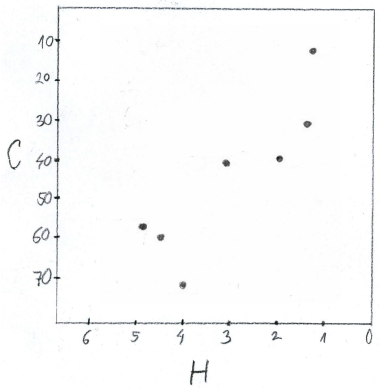
and a sequence of amino acid residues

.... 10-GLY 11-LYS 12-ILE 13-ARG 14-ALA....

1. Determine what amino acid corresponds to the frame
   (e.g. that the corresponding amino acid is ILE)
2. Determine the position of the frame in the sequence
   (e.g. that the corresponding residue is 12-ILE)

The plan

Given a CH frame



and a sequence of amino acid residues

.... 10-GLY 11-LYS 12-ILE 13-ARG 14-ALA....

1. Determine what amino acid corresponds to the frame
   (e.g. that the corresponding amino acid is ILE)
2. Determine the position of the frame in the sequence
   (e.g. that the corresponding residue is 12-ILE)

The amino acid recognition problem

Given a CH frame X_n and an amino acid A, estimate the probability that X_n corresponds to a residue of A.

Our approach: Rank *subsets* of peaks of X_n according to the probability of belonging to A.

We use machine learning models trained on tuples of peaks ... and are saved by **big data deep learning AI** HAHAHA!!

The amino acid recognition problem

Given a CH frame X_n and an amino acid A, estimate the probability that X_n corresponds to a residue of A.

Our approach: Rank *subsets* of peaks of X_n according to the probability of belonging to A.

We use machine learning models trained on tuples of peaks ... and are saved by **big data deep learning AI** HAHAHA!!

NO! Why?

- Only few proteins available (no big data).
- Even though we may use a quite large "surrogate" dataset of CH peaks to train our models, neural networks tend to overfit awfully (no deep learning so far).
- Too many possible subsets of peaks make naive ranking impossible.

The Data

We have

- NMR spectra for 12 proteins of hundreds of amino acid residues.
- I.e. too few tuples of peaks to train predictors for 20 amino acids for all subsets of peaks of CH frames!

The Data

We have

- NMR spectra for 12 proteins of hundreds of amino acid residues.
- I.e. too few tuples of peaks to train predictors for 20 amino acids for all subsets of peaks of CH frames!

Solution: Train models on data from Biological Magnetic Resonance Data Bank

- $\approx$ 8 millions of frequencies recorded by scientists since 2006.
- The data indicate what amino acid residue (and protein) the peaks came from.
- However, only well-formed tuples of peaks have been recorded!

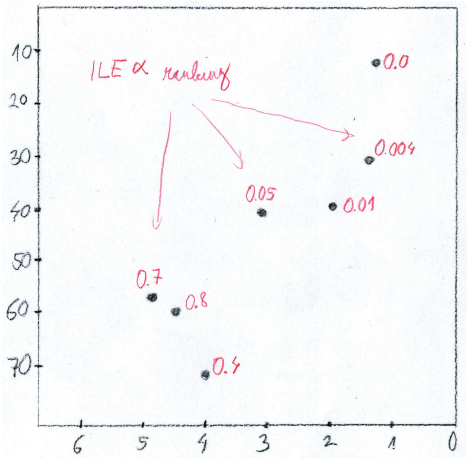
  Enormous number of possible "wrong" combinations of peaks makes construction of artificial data difficult.

Auxiliary models

On the BMRB data we train the following auxiliary models:

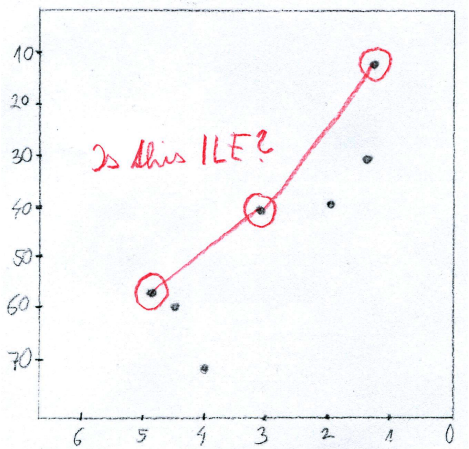
- Models ranking types of individual CH peaks.

  E.g. given a single pair of numbers (c, h), estimate the probability that they are chemical shifts of C_α and H_α of *ILE*.



- Models ranking amino acid types for tuples of peaks.

  E.g. given three pairs of numbers $(c_1, h_1), (c_2, h_2), (c_3, h_3)$ estimate the probability that they are peaks of *ILE*.



We use these models to estimate the probability that a given frame corresponds to a given amino acid.

Auxiliary models

- Ranking individual CH peaks: Histograms

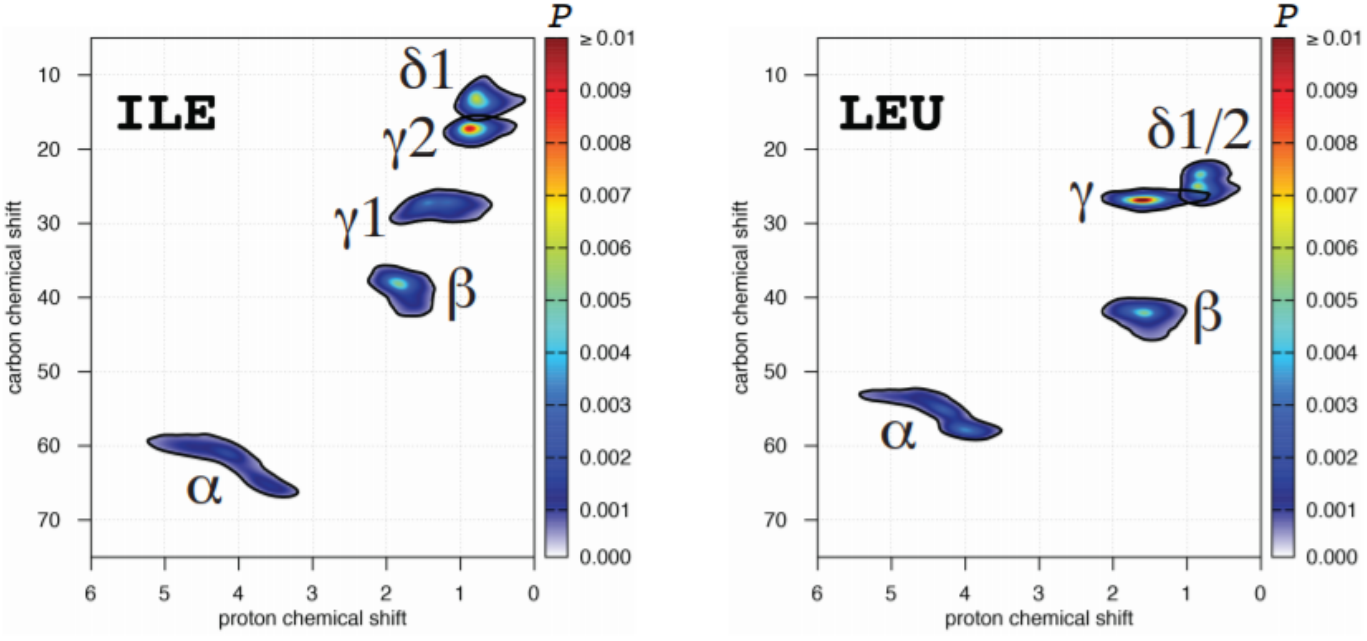
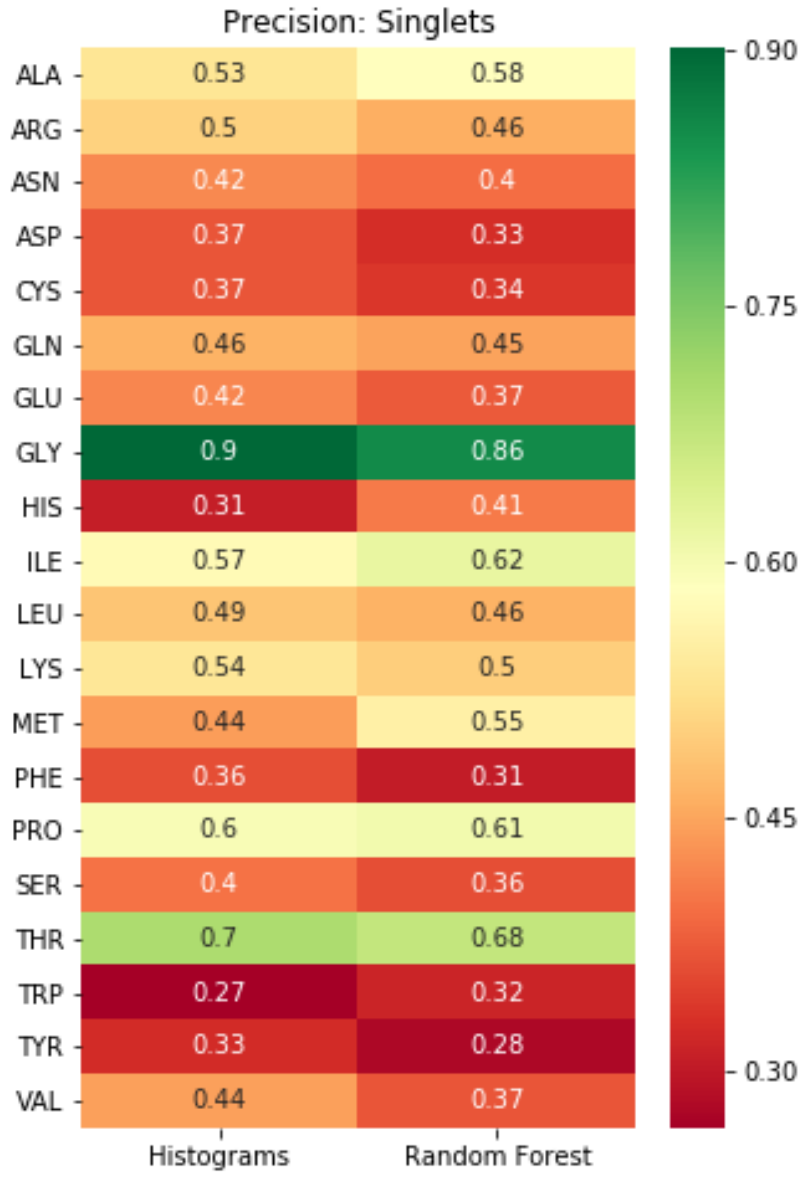
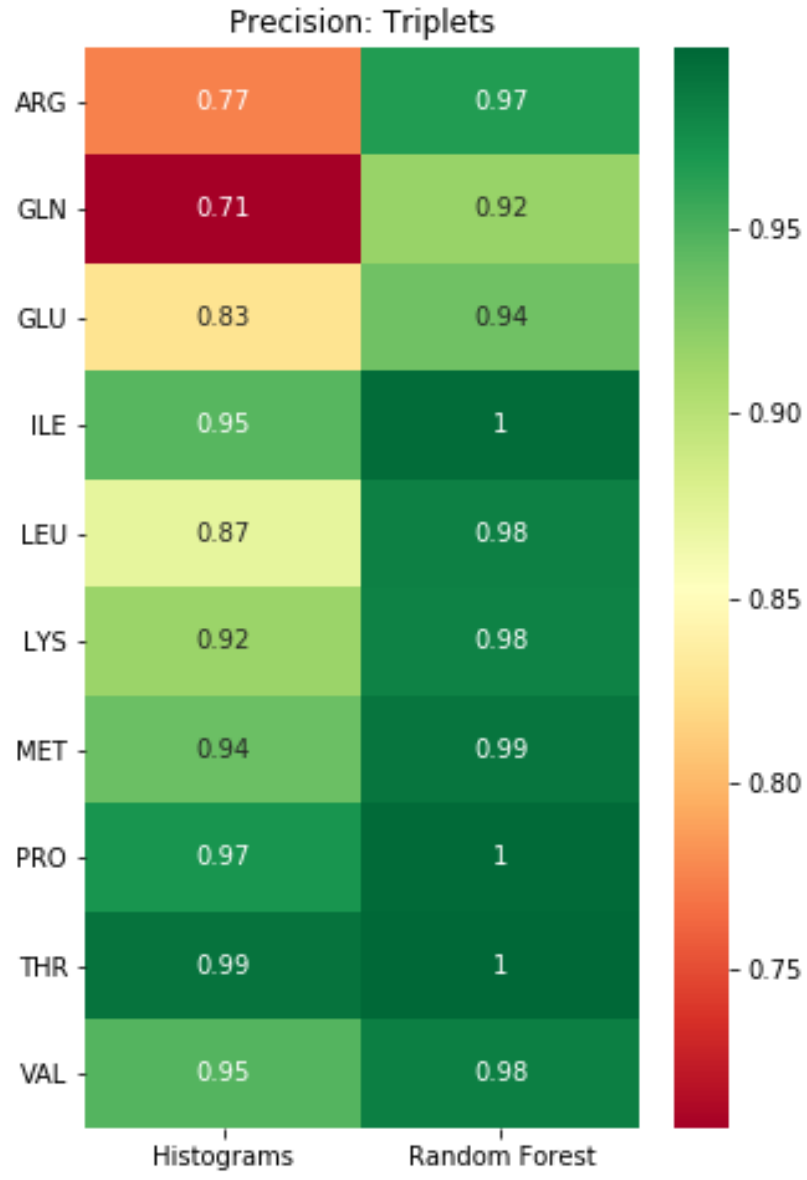


Image source: Automated NMR resonance assignments and structure determination using a minimal set of 4D spectra. Evangelidis et al. nature communications, 2018

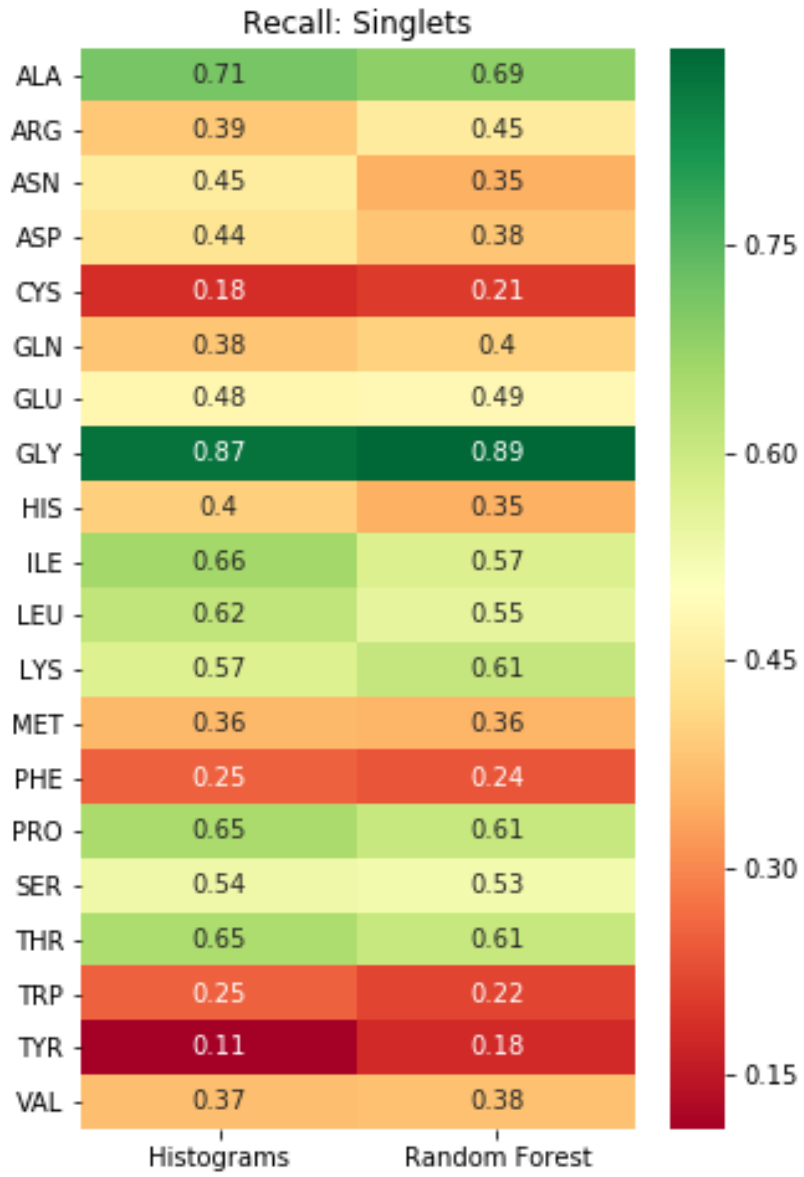
- Ranking amino acid types for tuples of peaks:
  - neural networks,
  - "naive" Bayes using the histograms,
  - **random forests**.

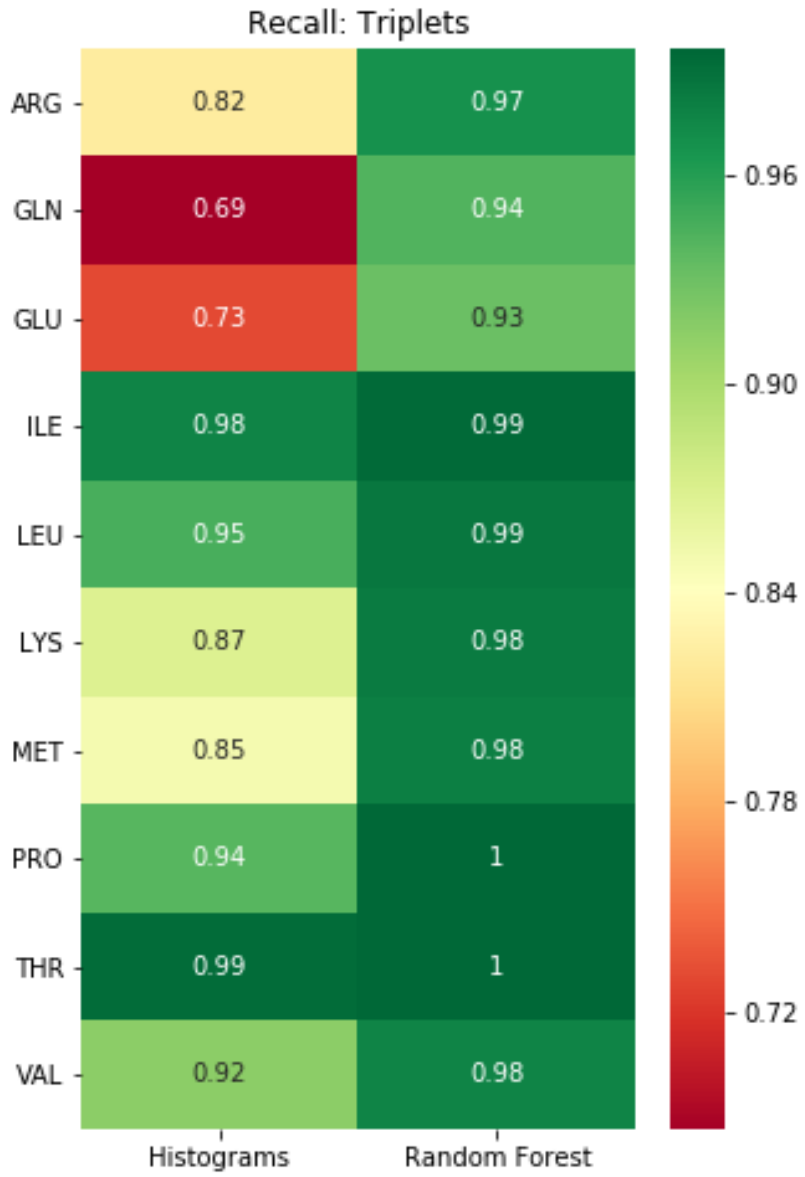
Auxiliary models – precision





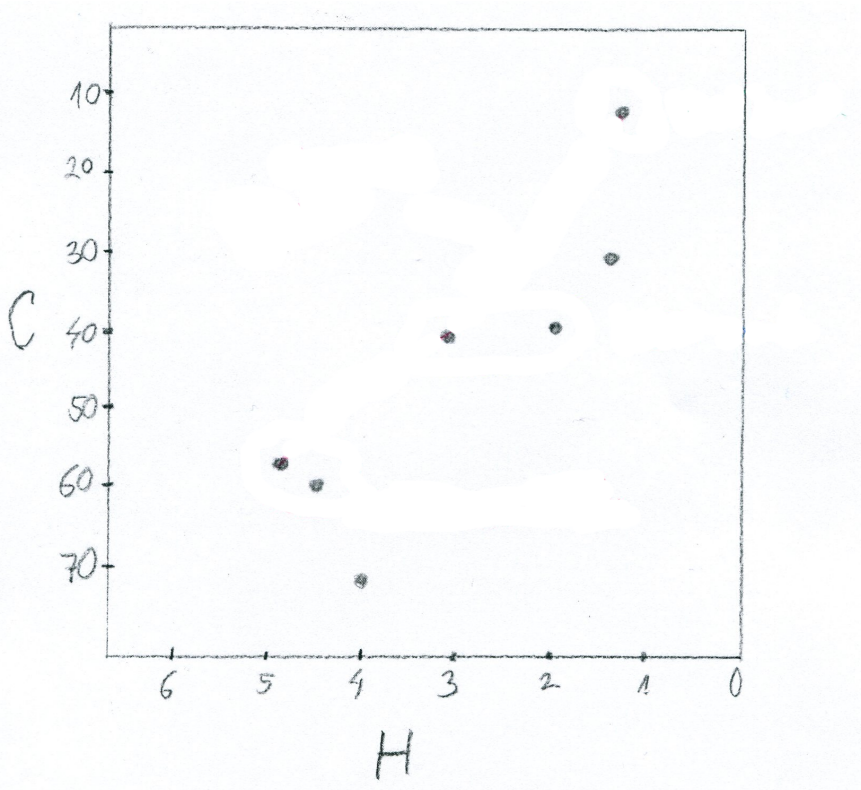
Auxiliary models – recall





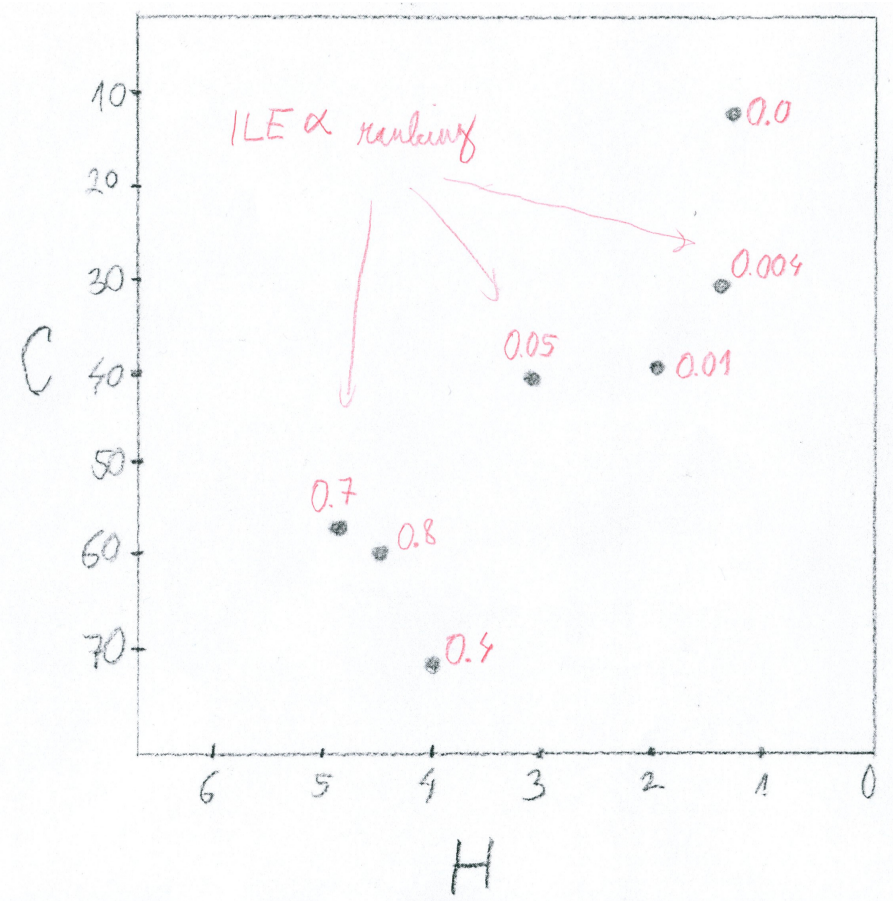
Amino acid recognition

Is ILE present in the following frame?



Amino acid recognition

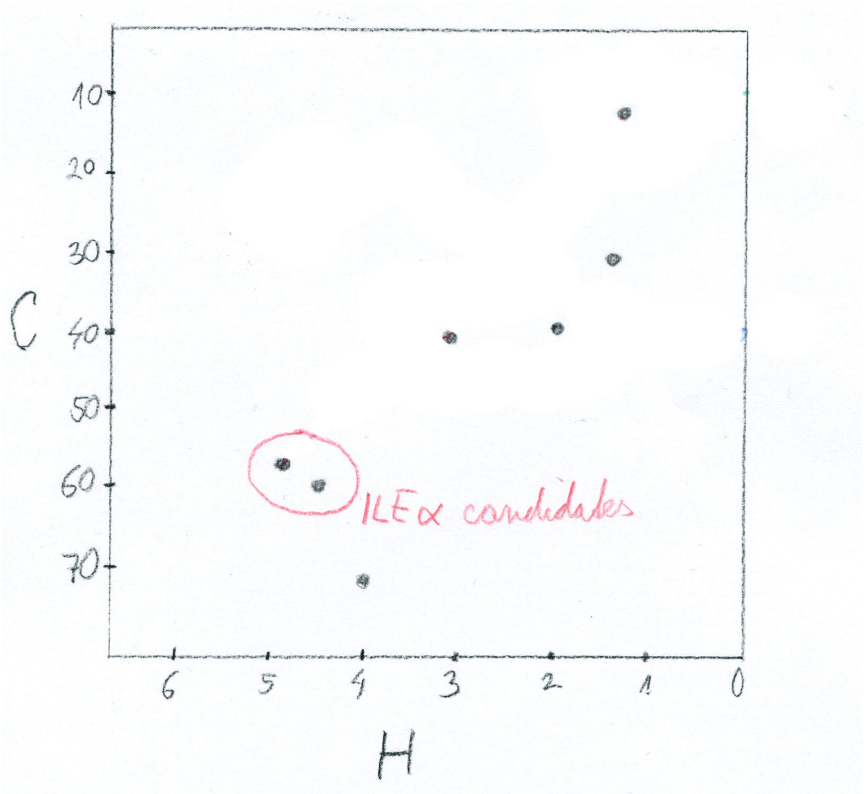
Is ILE present in the following frame?



Rank all peaks using a single peak classifier recognizing ILE-α peaks.

Amino acid recognition

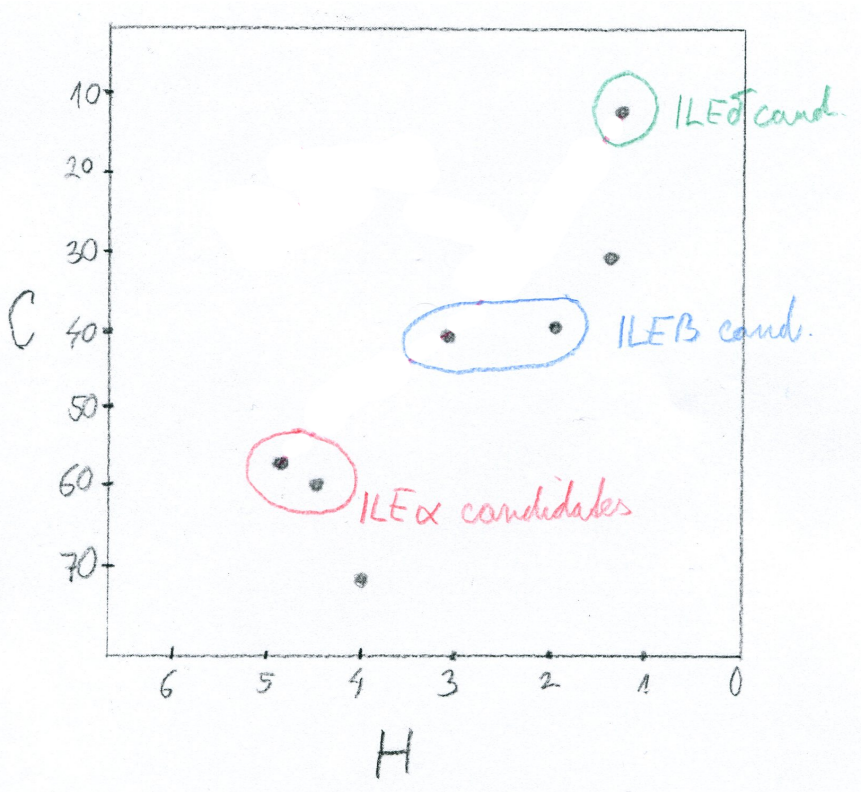
Is ILE present in the following frame?



Choose k highest ranking candidate ILE-α peaks (here $k = 2$) that make it above a threshold.

Amino acid recognition

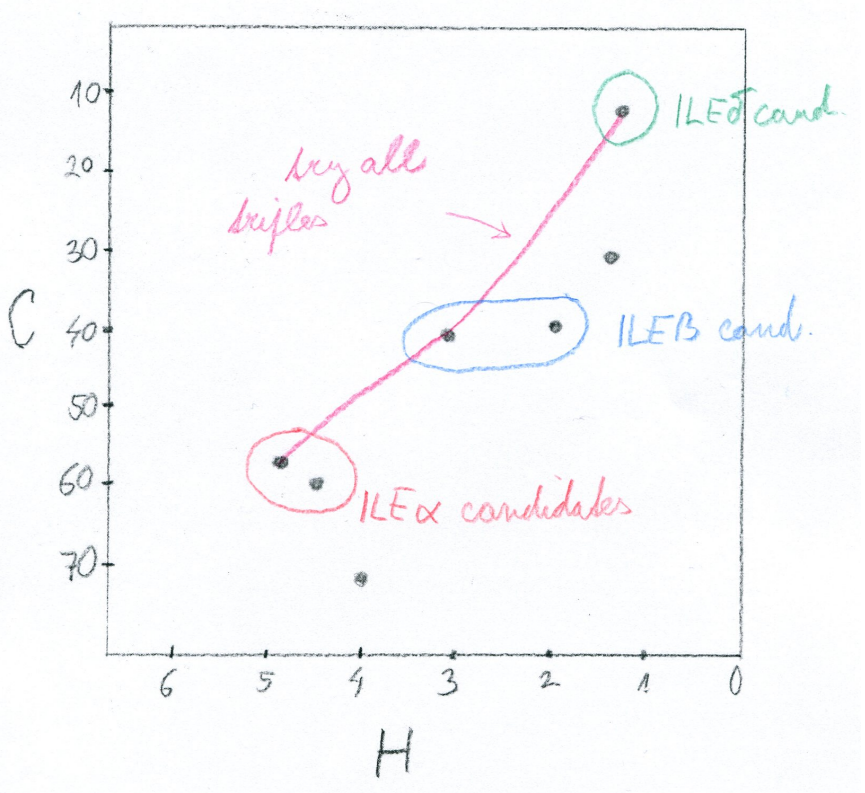
Is ILE present in the following frame?



Do the same with the remaining single peak models for ILE (i.e. *ILE–*β, *ILE–*γ, *ILE–*δ).

Amino acid recognition

Is ILE present in the following frame?



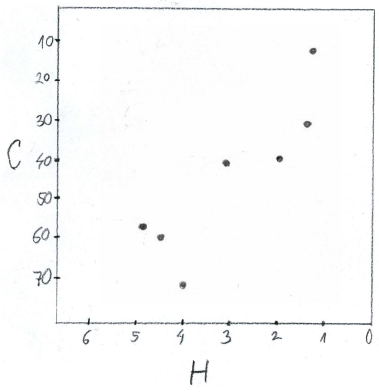
Consider all well-formed triples out of the candidate peaks.

Use the tuple based amino acid ranking model to rank each of the triples as being ILE.

The maximum of the ranks is the rank of ILE being in the frame.

The plan

Given a CH frame



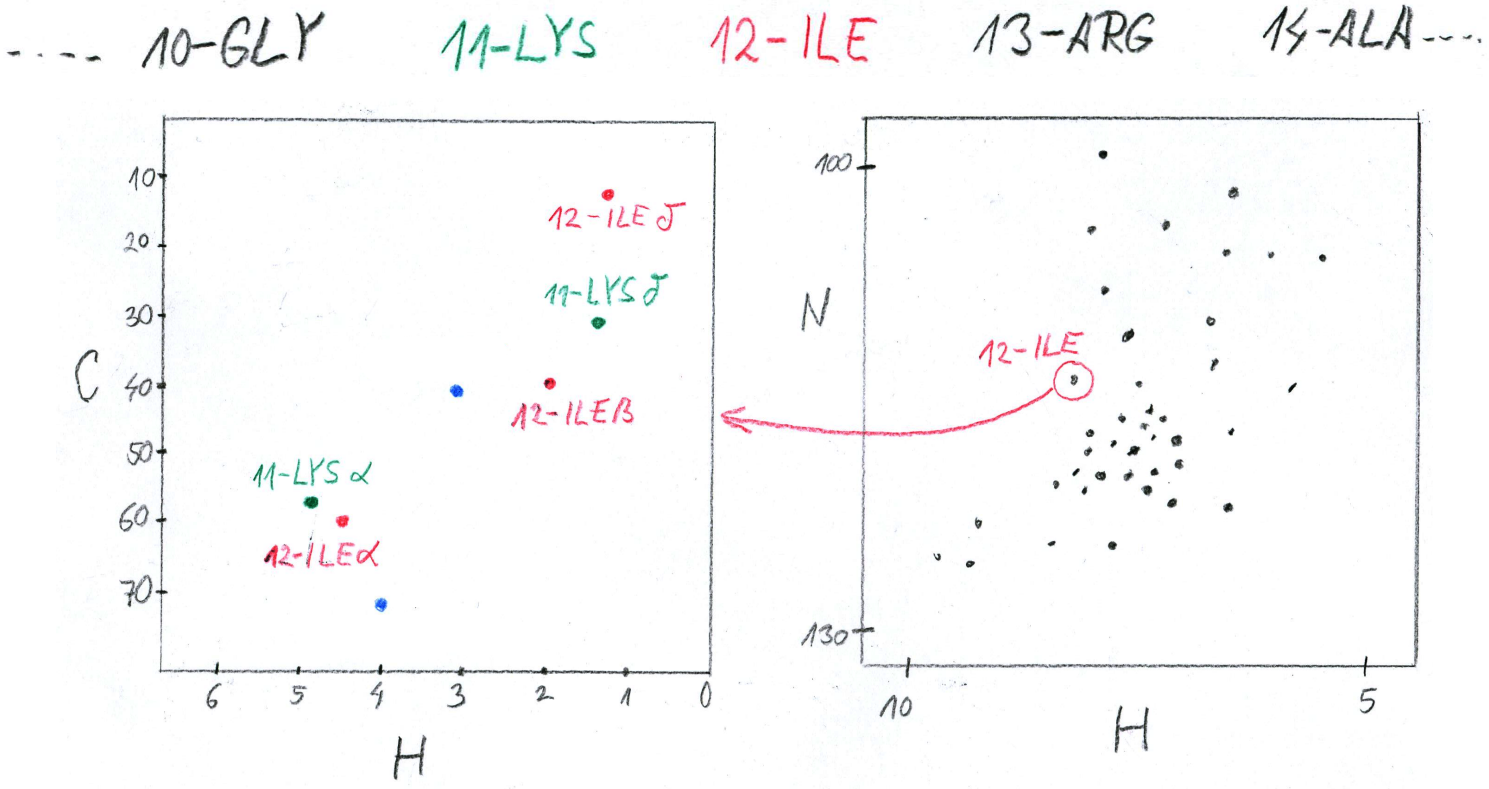
and a sequence of amino acid residues

.... 10-GLY 11-LYS 12-ILE 13-ARG 14-ALA....

1. Determine what amino acid corresponds to the frame
   (e.g. that the corresponding amino acid is ILE)
2. Determine the position of the frame in the sequence
   (e.g. that the corresponding residue is 12-ILE)

Sequential information

A sequence of amino acid residues

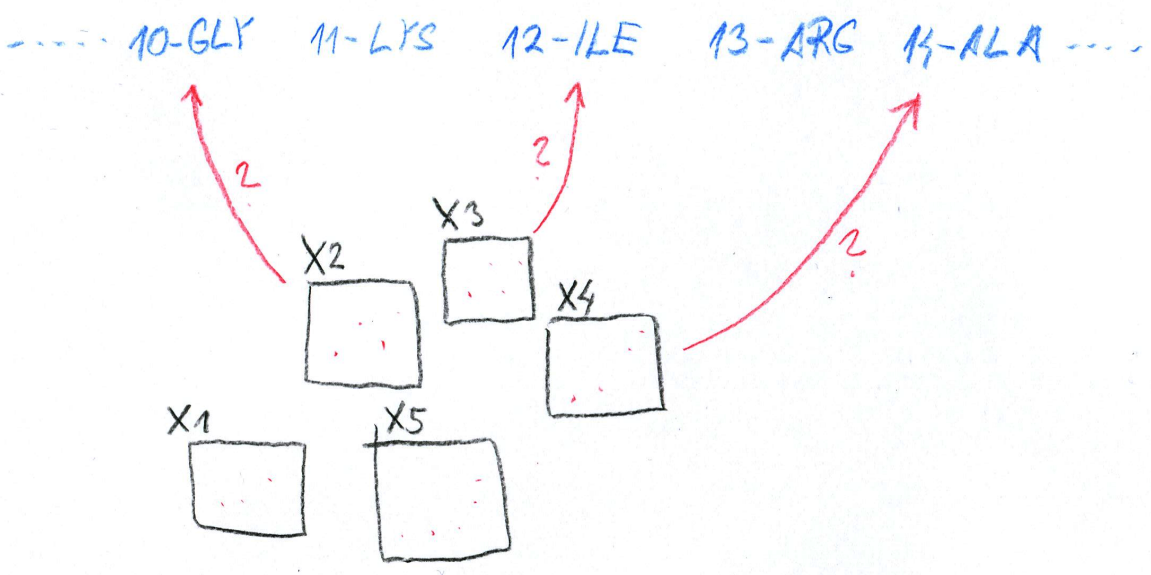


Crucial observation: Each CH frame typically contains

- peaks of the corresponding amino acid residue (here 12-ILE),
- peaks of the previous amino acid residue (here 11-LYS).

... so frames in sequence have higher number of *common peaks*.

Determine the position

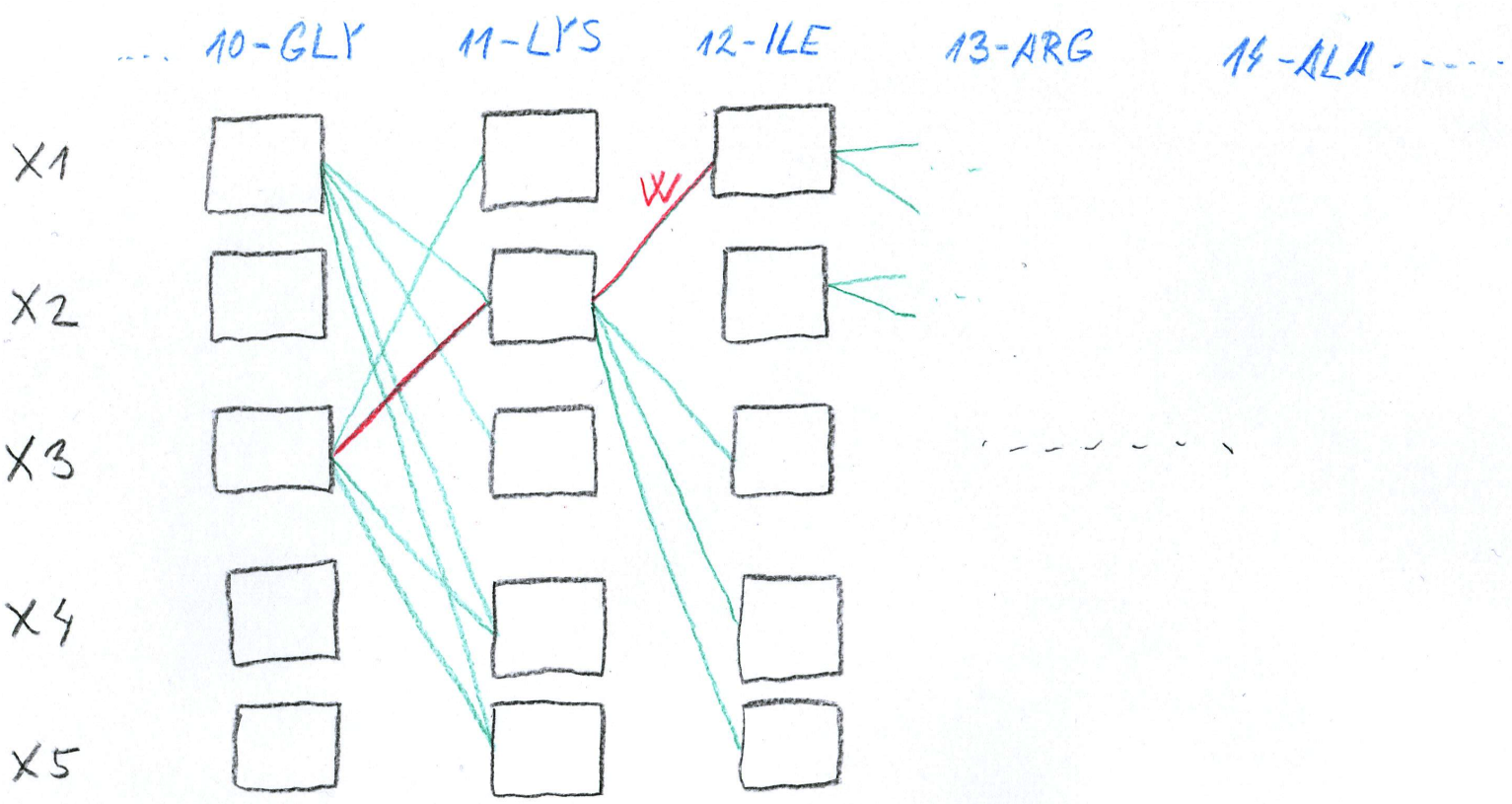


Available information for each CH frame X_n:

- What amino acids are *probably* present in X_n (the machine learning models explained earlier).
- What frames are *probably* next to X_n in the sequence (count common peaks, inspect intensity of peaks, etc.)

Use this to construct a weighted graph of frames aligned to the sequence of amino acid residues.

Connectivity graph



For example, the weight *w* contains information about:

- The number of common peaks between X_1 and X_2.
- The probability that the "right" amino acid of X_1 is *ILE*.

We search for a path maximizing the product of weights which visits each X_n at most once.

A simple path searching algorithm

Use a graph searching algorithm (e.g. Dijkstra) to find a path of maximum product weight.

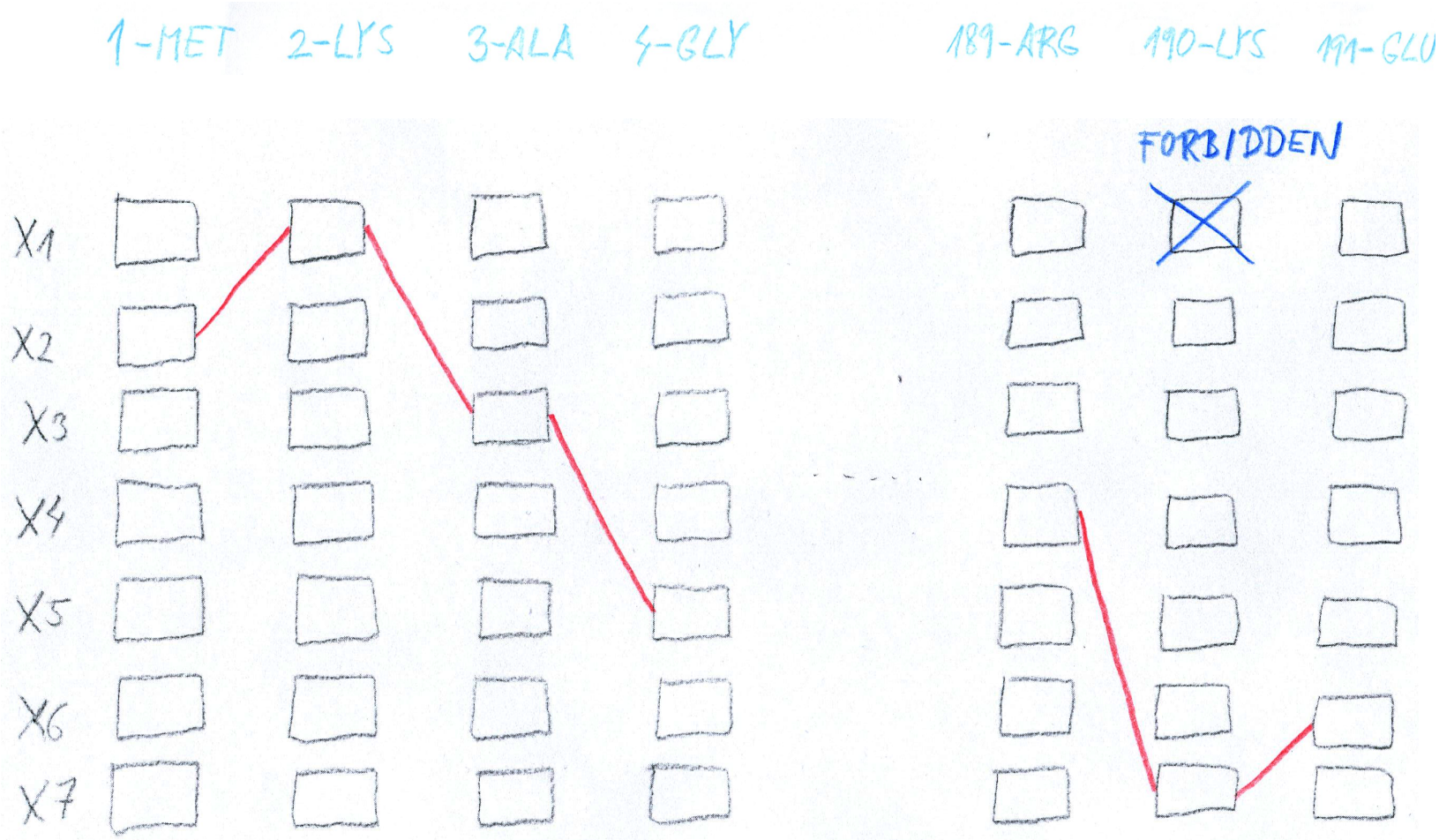
A simple path searching algorithm

Use a graph searching algorithm (e.g. Dijkstra) to find a path of maximum product weight.

A simple path searching algorithm

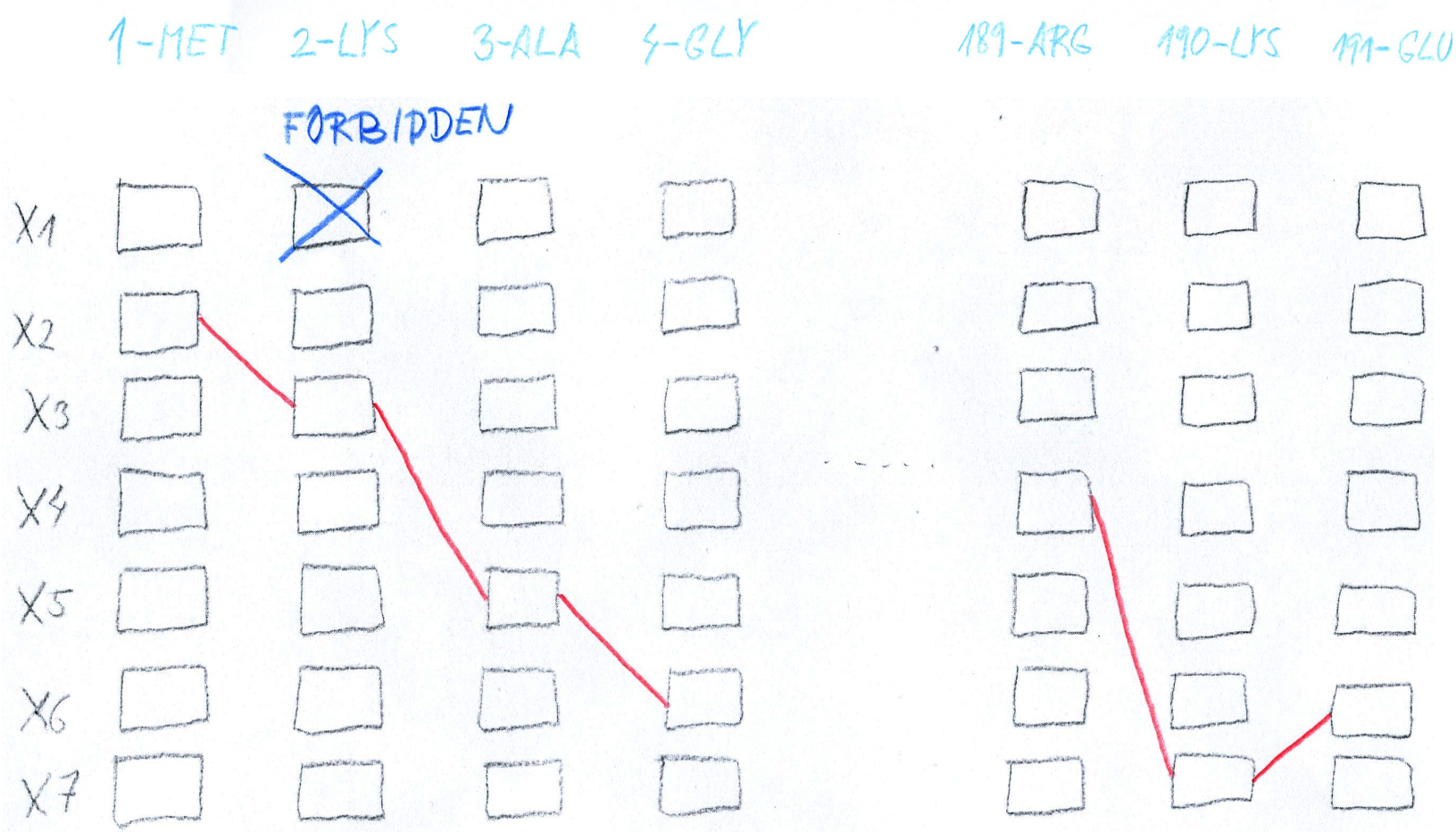
May contain duplicate visits to the same frame.

A simple path searching algorithm



Forbid each of the duplicate positions and apply the graph searching algorithm
(if more duplicate visits occur, continue recursively).

A simple path searching algorithm

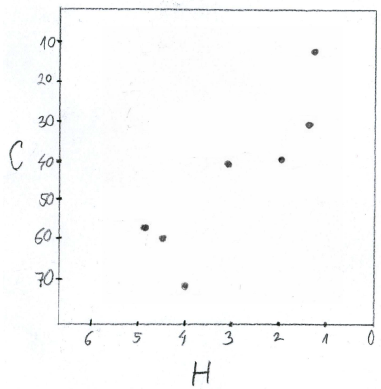


Forbid each of the duplicate positions and apply the graph searching algorithm
(if more duplicate visits occur, continue recursively).

Chose a path without duplicities of maximum product weight.

The plan

Given a CH frame



and a sequence of amino acid residues

.... 10-GLY 11-LYS 12-ILE 13-ARG 14-ALA ....

1. Determine what amino acid corresponds to the frame
   (e.g. that the corresponding amino acid is ILE)
2. Determine the position of the frame in the sequence
   (e.g. that the corresponding residue is 12-ILE)

DONE!

Summary

- The assignment problem for protein NMR spetroscopy.
- The algorithm employs machine learning techniques as well as graph searching algorithms.
- Work in progress.

Future work:

- Improve all parts of the algorithm.
- Apply to newly obtained NMR measurements.
- Integrate into a software for NMR analysis.