The influence of non-domain regions' composition on the activity of multi-domain protein kinases

Bachelor's thesis

Jan Hamalčík
Supervisor: doc. RNDr. Jiří Vondrášek, CSc.

Charles University
Faculty of Science

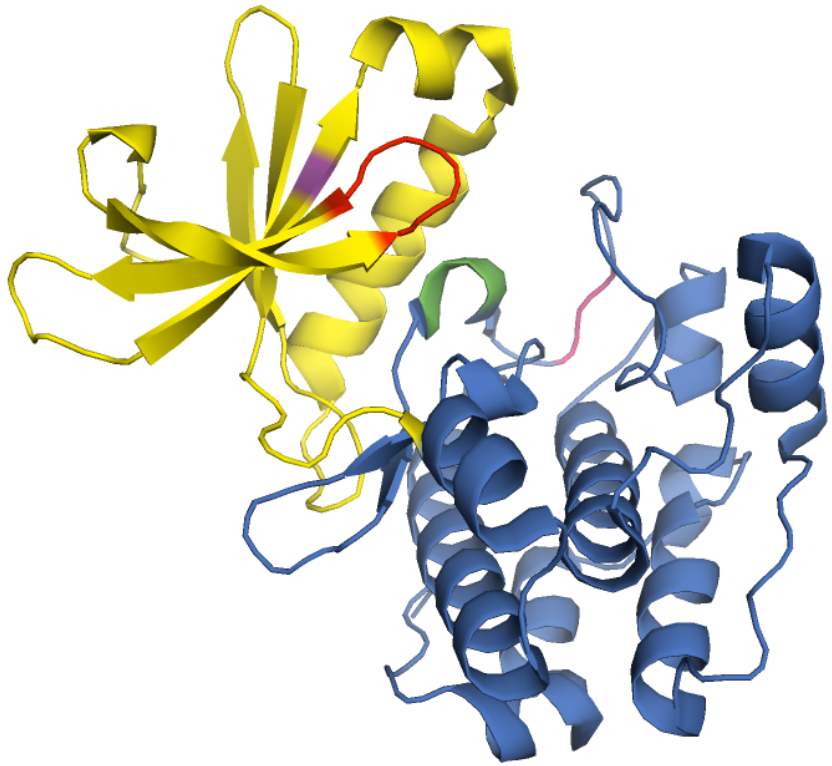
September 15th, 2020

Overview

- Hypothesis: General linker composition traits influence the function and specificity of the adjoint domains.
- Key consequence: Prediction of protein domain function based on the analysis of the non-domain regions.
- Method: On human two-domain proteins with one PK domain do:
  1. acquire physicochemical attributes of the linkers and average them over their sequences,
  2. cluster the proteins by the averaged physicochemical attributes,
  3. embed GO terms and EC numbers into the clustering.
- Result: No colocalization of like GO or EC terms associated with proteins with different architectures was observed.

Protein kinase domain

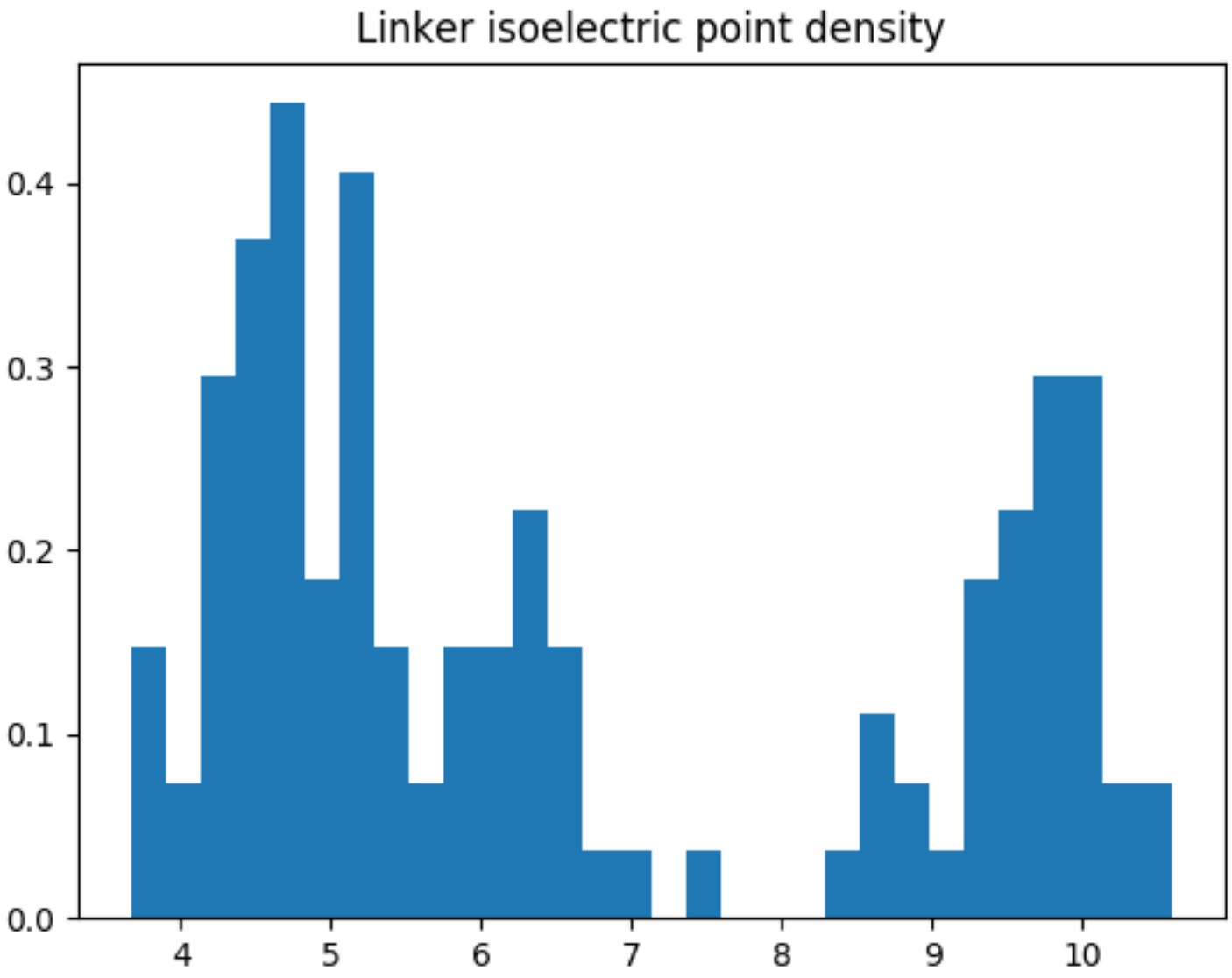
- Large and diverse family, involved in signal transduction.
- Phosphate group transfer from a phosphate donor onto a substrate.
- Bilobal structure with following conserved regions:
  - Gly-X-Gly-X-X-Gly motif
  - Invariant lysine
  - DFG motif
  - APE motif

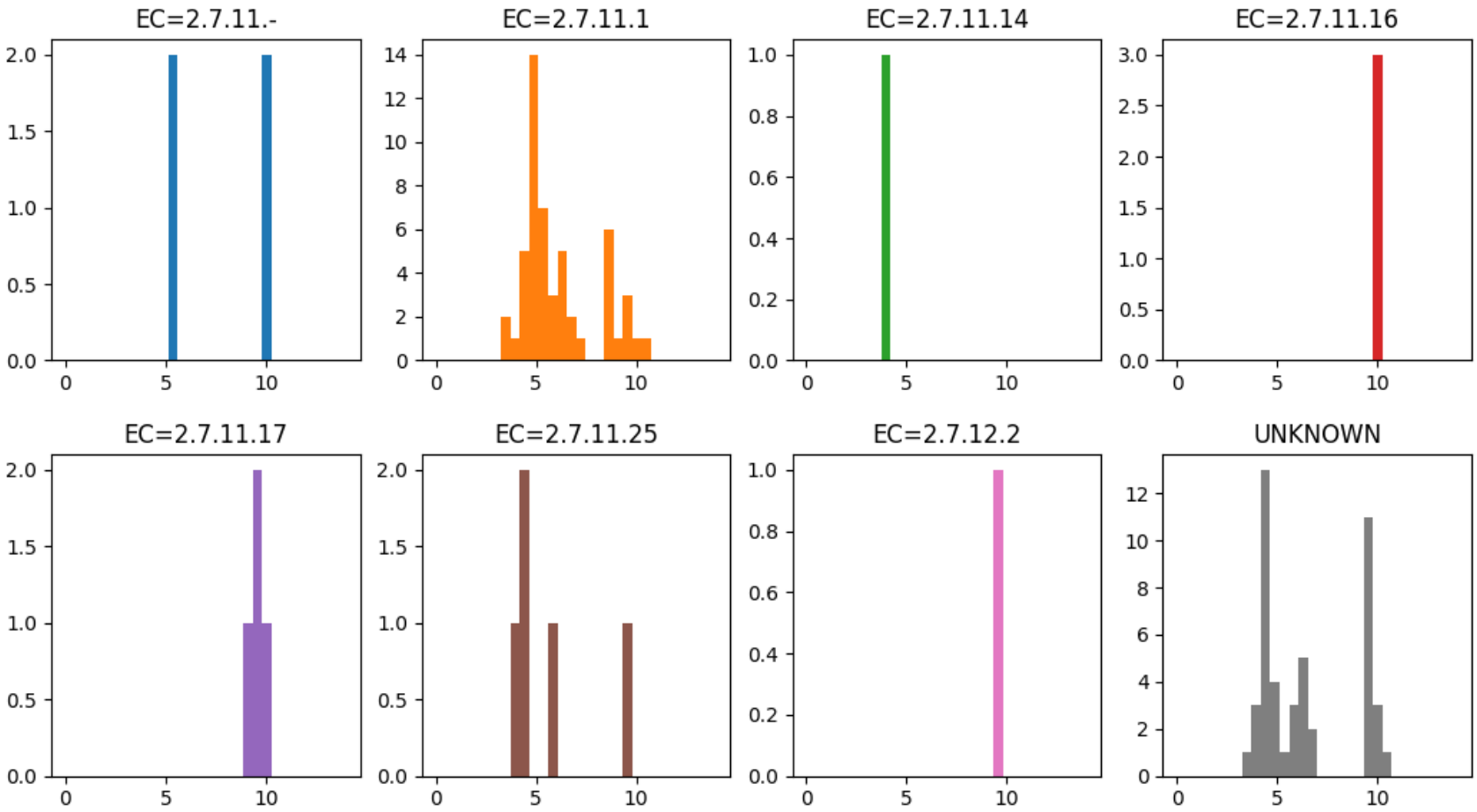


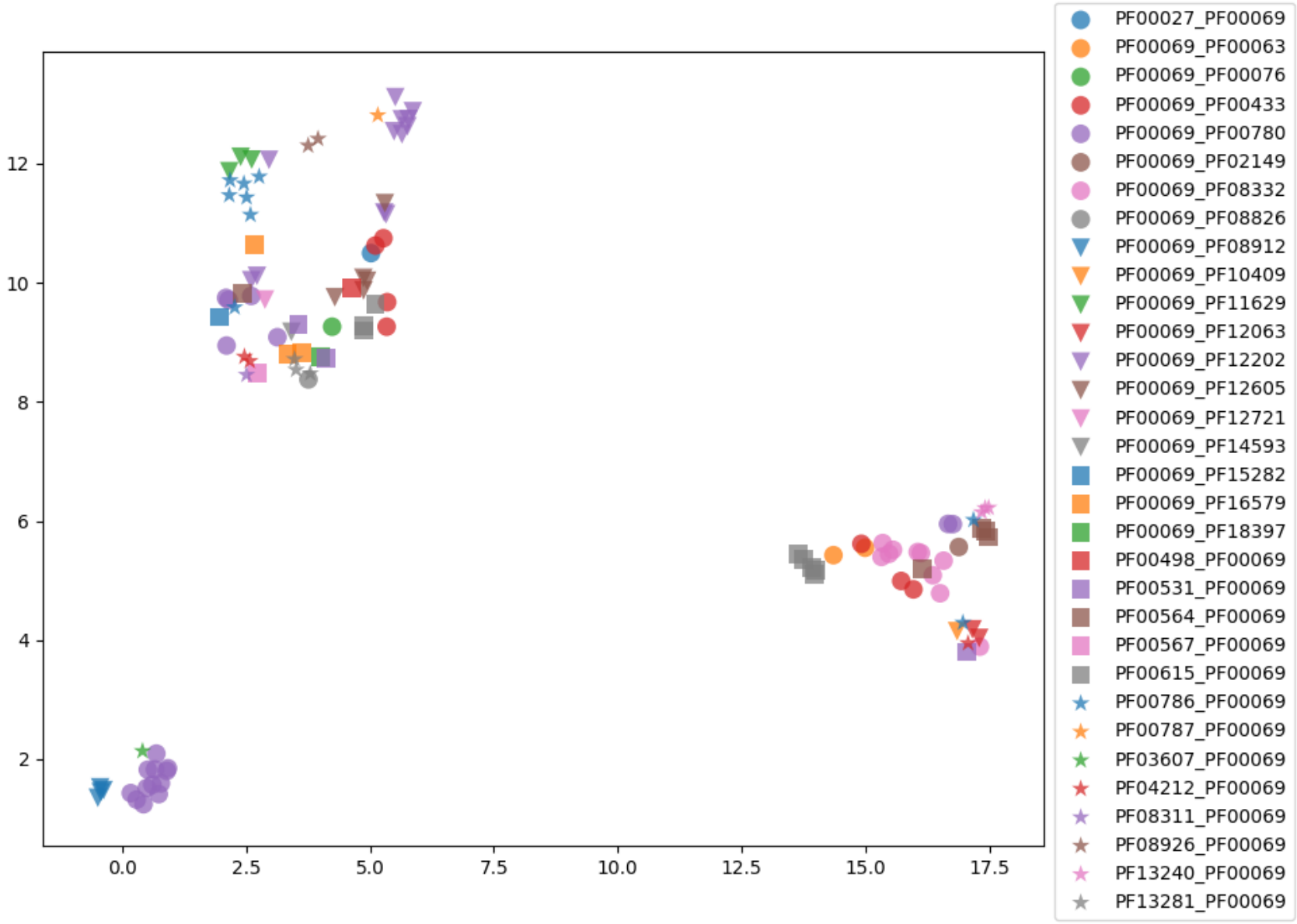
Dataset

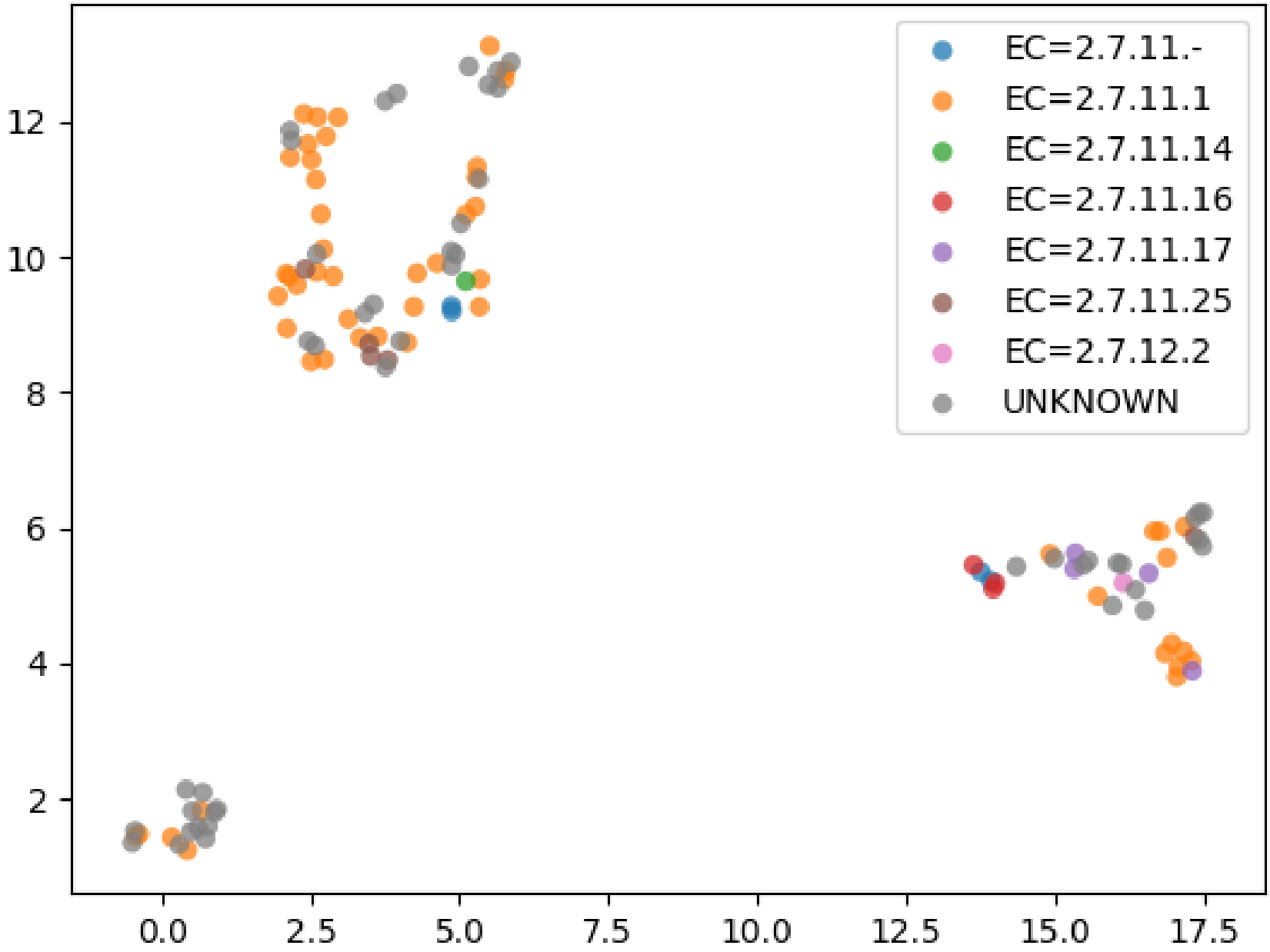
- 117 human two-domain proteins with one PK domain
- 32 different architectures
- Averaged physicochemical attributes of their inter-domain regions:
  - logarithm of the linker sequence length
  - isoelectric point
  - percentage of charged amino acids
  - GRAVY index

2D representation of the feature space performed by UMAP.


Linker isoelectric point density
0.4
0.3
0.2
0.1
0.0
4
5
6
7
8
9
10


EC=2.7.11.-
EC=2.7.11.1
EC=2.7.11.14
EC=2.7.11.16
EC=2.7.11.17
EC=2.7.11.25
EC=2.7.12.2
UNKNOWN


12
10
8
6
4
2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
PF00027_PF00069
PF00069_PF00063
PF00069_PF00076
PF00069_PF00433
PF00069_PF00780
PF00069_PF02149
PF00069_PF08332
PF00069_PF08826
PF00069_PF08912
PF00069_PF10409
PF00069_PF11629
PF00069_PF12063
PF00069_PF12202
PF00069_PF12605
PF00069_PF12721
PF00069_PF14593
PF00069_PF15282
PF00069_PF16579
PF00069_PF18397
PF00498_PF00069
PF00531_PF00069
PF00564_PF00069
PF00567_PF00069
PF00615_PF00069
PF00786_PF00069
PF00787_PF00069
PF03607_PF00069
PF04212_PF00069
PF08311_PF00069
PF08926_PF00069
PF13240_PF00069
PF13281_PF00069


EC=2.7.11.-
EC=2.7.11.1
EC=2.7.11.14
EC=2.7.11.16
EC=2.7.11.17
EC=2.7.11.25
EC=2.7.12.2
UNKNOWN
12
10
8
6
4
2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

Conclusion

- No general influence of the linkers' composition on the overall protein function observed.
- Problems:
  - 3.65625 proteins per architecture on average
  - GO and EC are too general and too precise, respectively
  - Observed linker types are determined by UMAP parametrization