Computational Biology and Data Mining

TH BINGEN
University of Applied Sciences

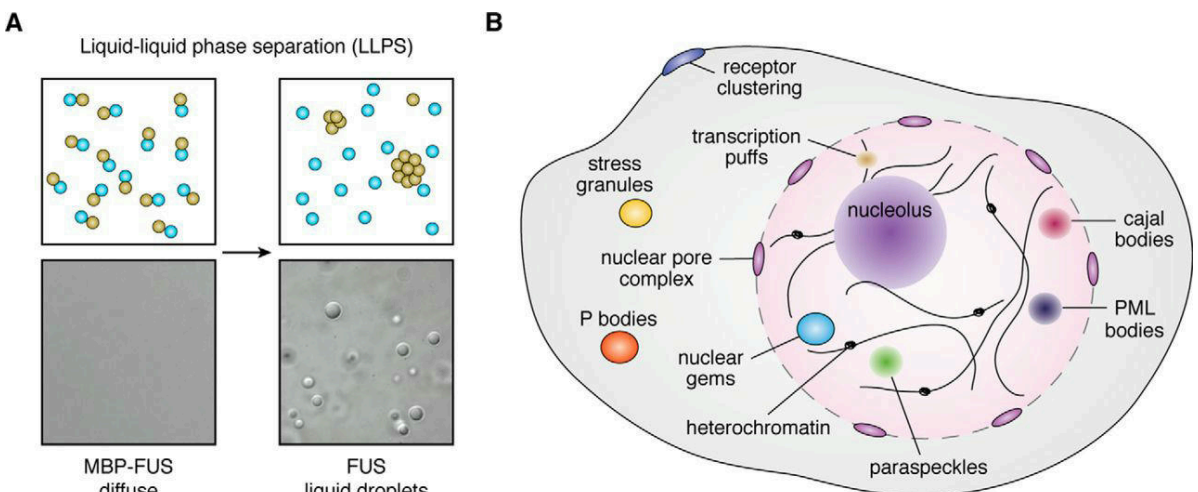
Prediction of Liquid-Liquid Phase Separation of Proteins using Neural Networks

Bachelorarbeit Robin Ender

Datum: 07.08.2025

Liquid-Liquid Phase Separation (LLPS)

- Bildung von Membranlosen Organellen
- Pathologisches Wirken: Demenz, Krebs, Infektionen

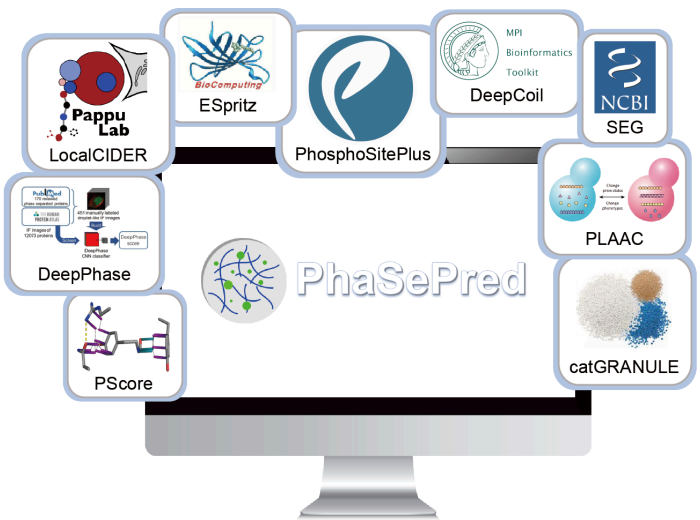


[1]E. Gomes and J. Shorter, "The molecular language of membraneless organelles," *Journal of Biological Chemistry*, vol. 294, no. 18, pp. 7115–7127, May 2019, doi: 10.1074/jbc.TM118.001192.

[2]B. Wang *et al.*, "Liquid–liquid phase separation in human health and diseases," *Signal Transduction and Targeted Therapy*, vol. 6, no. 1, p. 290, Aug. 2021, doi: 10.1038/s41392-021-00678-1.

LLPS Predictor

- Erste Generation: Hidden Markov Modelle, Formel basiert
- Zweite Generation: Machine Learning (die skalare Werte)
- Einteilung der LLPS Proteine nach Selbst-Assemblierend | Partner-Abhängig und Ungeordnet | Geordnet



[1] Z. Chen *et al.*, "Screening membraneless organelle participants with machine-learning models that integrate multimodal features," *Proceedings of the National Academy of Sciences*, vol. 119, no. 24, p. e2115369119, Jun. 2022, doi: 10.1073/pnas.2115369119.

Hypothesen

- Neuronale Netze könnten besser geeignet für LLPS Prediction sein, als heutige tools, weil sie:
  - Reihenfolge und Anordnung von Aminosäuren berücksichtigen, dadurch komplexere Zusammenhänge verstehen können
  - Feature Engineering reduzieren
- Input: Block Decomposition | Sequenz
- Ein neuer Datensatz des PPMC-lab, der speziell für die Entwicklung von LLPS Predictors entwickelt wurde, sollte auch getestet werden

Datensatz und Evaluation

Hauptdatensatz: PSPire

- Etwa 10.000 negativ gelabelte Proteine und etwa 500 positiv gelabelte Proteine

Evaluation durch:

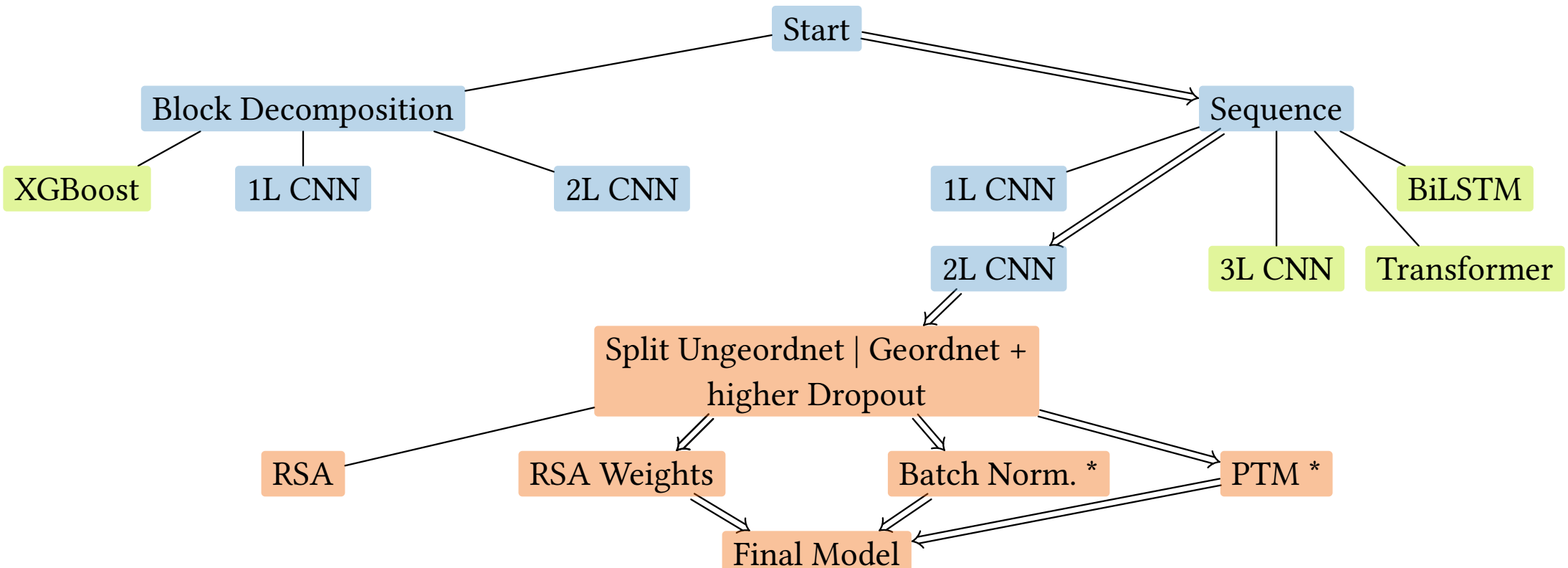
- Vergleich der ROCAUC und PRAUC Werte zu denen aus anderen Studien
  - Vor allem die PRAUC ist eine wichtige Metrik für unausgeglichene Datensätze

Modellfindung I

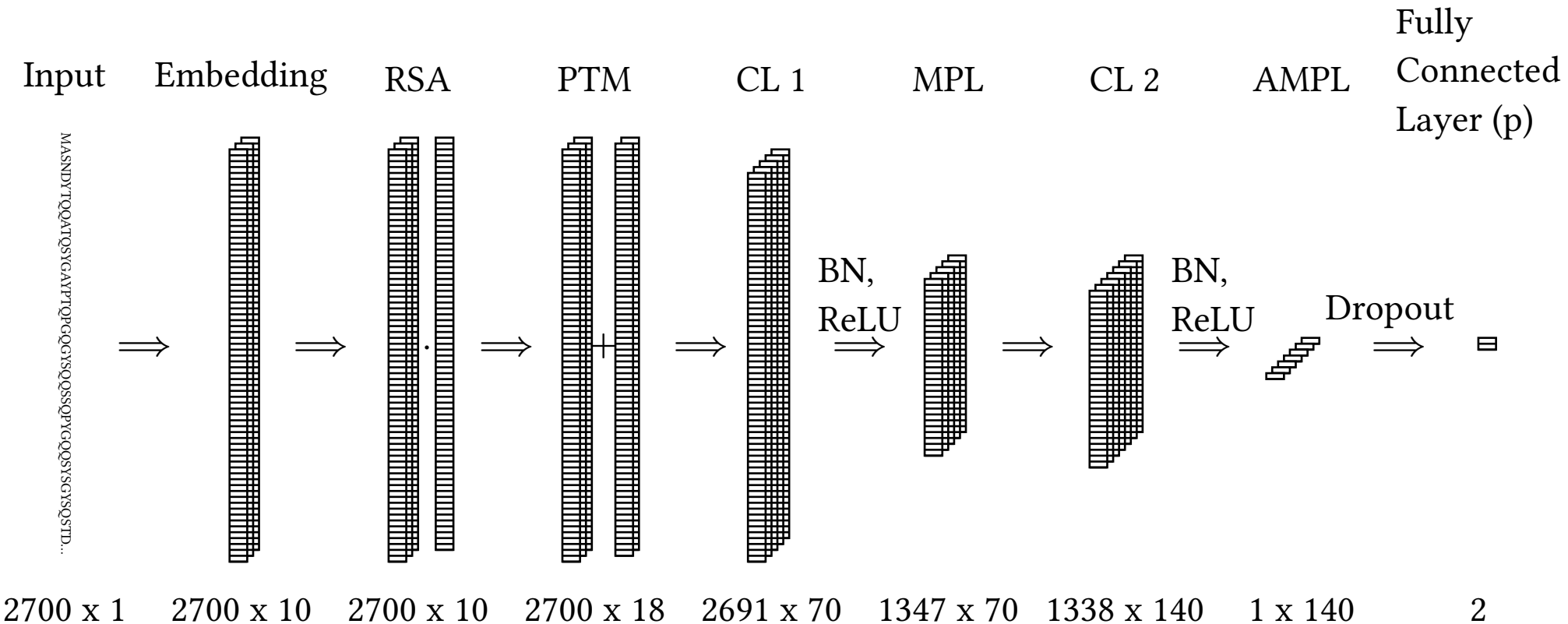
Modellfindung über drei Phasen:

1. Vergleich Block Decomposition gegen Sequenz +
   Test ob Neuronale Netze lernfähig sind
2. Testen weiterer komplexerer Modelle
3. Optimierung des besten Modells und Zugabe weiterer biologischer
   Information an die Modelle

Modellfindung II



Finales Modell



Ergebnisse PSPire Datensatz

Vergleich mit PSPire [1] und PdPS[2]:

AUC	IDR			non-IDR		
	Final Model	PSPire	PdPS	Final Model	PSPire	PdPS
ROC	0.80	0.86	0.84	0.88	0.84	0.68
PRC	0.42	0.51	0.42	0.25	0.24	0.08

Vergleich mit PSPire und PdPS auf MLO Daten:

- ähnlich wie PSPire Datensatz

---

[1] S. Hou, J. Hu, Z. Yu, D. Li, C. Liu, and Y. Zhang, "Machine learning predictor PSPire screens for phase-separating proteins lacking intrinsically disordered regions," *Nature Communications*, vol. 15, no. 1, p. 2147, Mar. 2024, doi: 10.1038/s41467-024-46445-y.

[2] Z. Chen *et al.*, "Screening membraneless organelle participants with machine-learning models that integrate multimodal features," *Proceedings of the National Academy of Sciences*, vol. 119, no. 24, p. e2115369119, Jun. 2022, doi: 10.1073/pnas.2115369119.

Ergebnisse catGranule / PPMC-lab Datensatz

Vergleich catGranule 2.0 [1]:

- Leicht bessere Performance als die dort beschriebenen Tools
  - catGranule 2.0: 0.76 ROCAUC
  - non-IDP Modell: 0.80 ROCAUC

PPMC-lab Datensatz:

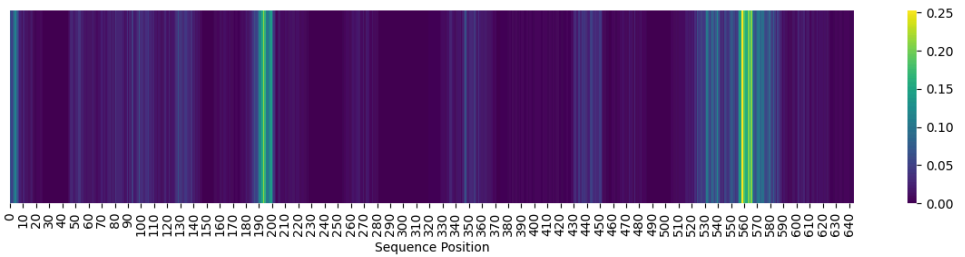
- Evaluation auf MLO Daten zeigte deutlich schlechtere Performance

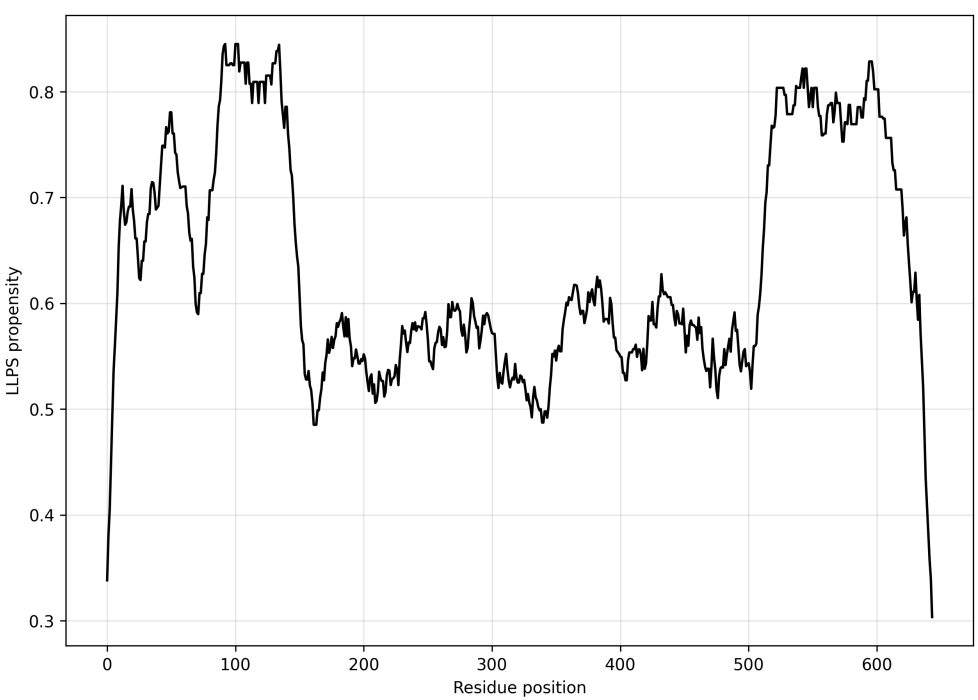
[1] M. Monti *et al.*, "catGRANULE 2.0: accurate predictions of liquid-liquid phase separating proteins at single amino acid resolution," *Genome Biology*, vol. 26, no. 1, p. 33, Feb. 2025, doi: 10.1186/s13059-025-03497-7.

Ergebnisse Saliency

Saliency map und LLPS

Profil von Protein P04264





Fazit

Fazit

- Block Decomposition als Input nicht besser als die Sequenz

Fazit

- Block Decomposition als Input nicht besser als die Sequenz
- Trotz relativ kleinen Datensätzen haben die CNNs bereits vergleichbare Leistung zu anderen State of the Art Tools gezeigt -> CNNs haben ggf. das Potenzial bei besserer Datenlage jetzige Tools abzulösen (mehr und besser annotiert - PTMs)

Fazit

- Block Decomposition als Input nicht besser als die Sequenz
- Trotz relativ kleinen Datensätzen haben die CNNs bereits vergleichbare Leistung zu anderen State of the Art Tools gezeigt -> CNNs haben ggf. das Potenzial bei besserer Datenlage jetzige Tools abzulösen (mehr und besser annotiert - PTMs)
- Saliency ermöglicht Einblicke in Entscheidungsfindung, interpretierbar wie LLPS Profil anderer Tools

Outlook

Outlook

- Saliency Maps weiter analysieren

Outlook

- Saliency Maps weiter analysieren
- Weitere Optimierung des Modells durch systematische Anpassung der Parameter und five-fold Validation

Outlook

- Saliency Maps weiter analysieren
- Weitere Optimierung des Modells durch systematische Anpassung der Parameter und five-fold Validation
- Experimente mit grösserem Datensatz wiederholen

Outlook

- Saliency Maps weiter analysieren
- Weitere Optimierung des Modells durch systematische Anpassung der Parameter und five-fold Validation
- Experimente mit grösserem Datensatz wiederholen
- Ein fertiges Tool (Web Anwendung / CLI-Tool) entwickeln

Vielen Dank!

Fragen?