Multi-Objective Active Learning for Nanobody Development

Katharina Dost *[1], **Klara Kropivšek*[2]**, Christian L. Camacho Villalón[1], Sašo Džeroski[1], and Ario de Marco[2]

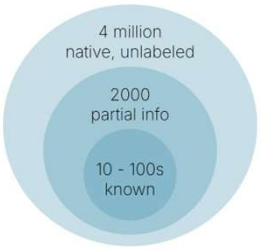
[1] Jožef Stefan Institute, Slovenia
[2] Laboratory for Environment and Life Sciences, University of Nova Gorica, Slovenia

Motivation

Nanobodies:
- small "**keys**" that match specific sites on cells, proteins
- can bind to **block or modulate specific functions**
- can **carry payload** (e.g., drug, fluorescent marker)

Advantages:
- **Small size** → reach hidden epitopes that classical antibodies cannot reach
- **Stable & robust** → more tolerant to harsh conditions
- **Versatile** → can be engineered for therapy, diagnostics, research

Dataset



Small dataset (10 - 100s): in-house nanobodies with measured yields & developability (high-quality labels).
Medium dataset (~2,000): nanobodies with partial annotations (literature/structural).
Largest dataset (4 million): non-redundant native repertoire sequences (**no experimental data, unlabeled**).

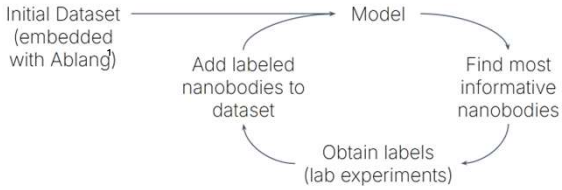
Research Question

Obtaining experimental data is **costly**. A great ML model could guide discovery as it can point to potential candidates for a task, ensuring **developability** in the wet lab. A great model needs **data** to learn from.

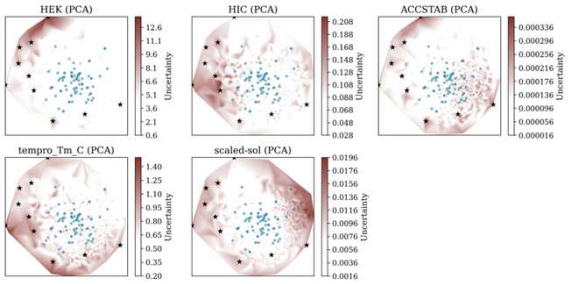
Which experiments (= labeled data) would help to train such a model?

Active Learning

Big-Picture Idea



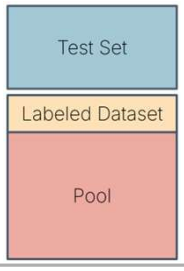
Uncertainty



Uncertainty Map highlighting areas with uncertain predictions per target in Ablang embedding space (dimensionality reduced with PCA). Blue = initial nanobodies; black = selected batch of nanobodies.

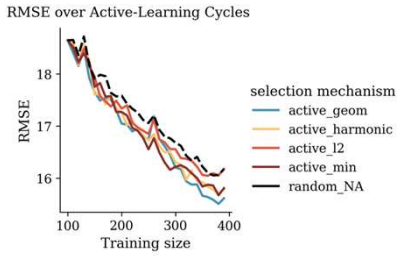
Active Learning Cycle

In each AL cycle, to select **batch of nanobodies (e.g., 10):**
- filter based on **constraints** for yield, thermal stability, solubility (as predicted)
- drop **outliers**
- select **top X** nanobodies that **maximize uncertainty** in predicting yield, polyreactivity, accelerated & thermal stability, solubility
- greedily select **diverse subset**



Simulation

- Simulation on **3000 nanobodies** from our native library
- **Labels** were obtained using Abpred[2], TEMPRO[3] and Protein-Sol[4]
- Per-target models with per-target uncertainties; different **aggregation strategies**



Conclusion

Summary:
- Multi-objective active nanobody selection strategy for better property predictions

Next Steps:
- Model catering to distribution shift
- Refined multi-objective optimization techniques
- Tests on complete native library
- Real AL cycle + tags

Acknowledgements. Dost, Kropivšek and Camacho are supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH, co-funded under the grant agreement No. 101081355. The SMASH project is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund. Džeroski and de Marco are supported by the Slovenian Research and Innovation Agency (under grants P2-0103, GC-0001, and P3-0428, N4-0282, N4-0325, J4-50144, respectively).

References

[1] https://github.com/oxpig/AbLang2
[2] https://github.com/maxhebditch/abpred
[3] https://github.com/Jerome-Alvarez/TEMPRO
[4] https://protein-sol.manchester.ac.uk/

Scan me to connect!

1995-2025 UNIVERZA V NOVI GORICI · Jožef Stefan Institute · SMASH