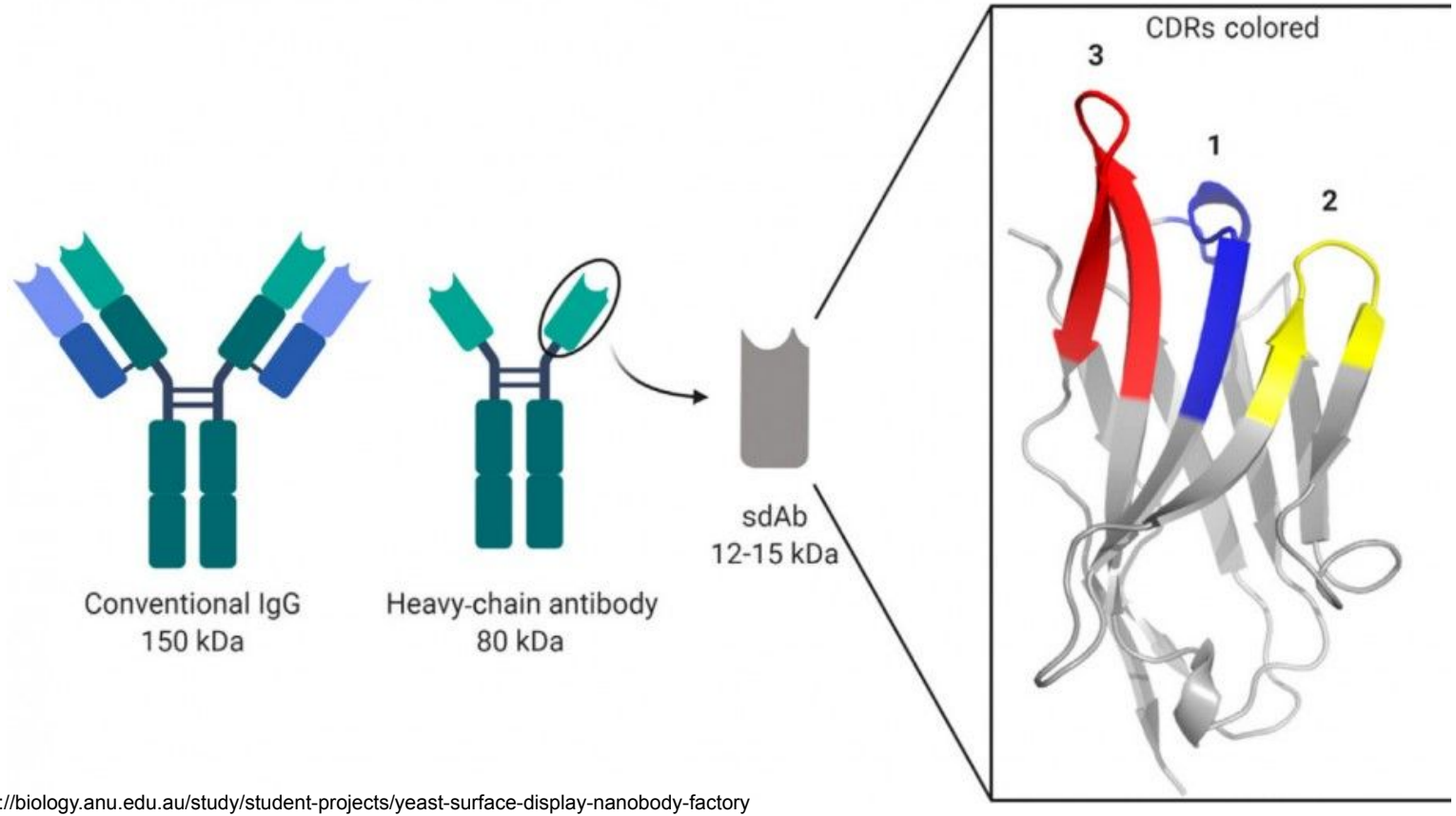
Model validation strategy is crucial to properly estimate the real-life applicability of machine learning models

Earth

Nanobodies are small antibodies



https://biology.anu.edu.au/study/student-projects/yeast-surface-display-nanobody-factory

Can we find alternative nanobodies that are more stable than wild type?

Suggested pipeline:

...ISKHATD**Y**GF...

Suggested pipeline: find functional mutated sequences

...ISKHATD**Y**GF...

nanoBERT

(Hadsund, J.T. et al. 2024)

...ISKHATD**F**GF...

Mutation in CDR3

Suggested pipeline: predict 3D structure

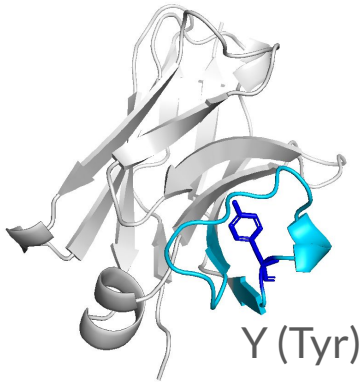
...ISKHATD**Y**GF...

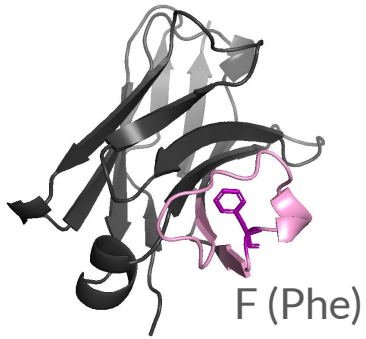
nanoBERT

(Hadsund, J.T. et al. 2024)

...ISKHATD**F**GF...

Mutation in CDR3





AlphaFold2 structure prediction

(Jumper, J. et al. 2021. Nature)

Suggested pipeline: predict melting temperature

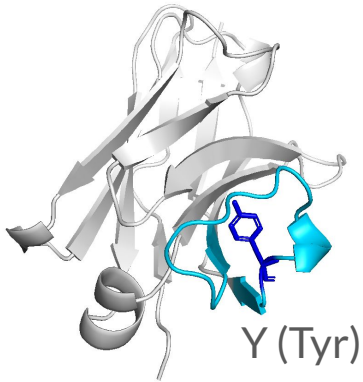
...ISKHATD**Y**GF...

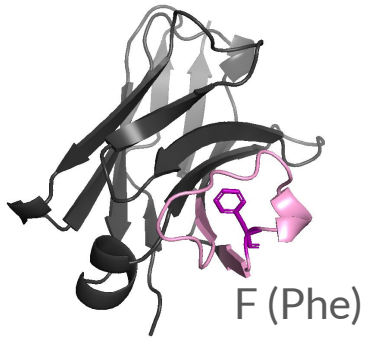
nanoBERT

(Hadsund, J.T. et al. 2024)

...ISKHATD**F**GF...

Mutation in CDR3





AlphaFold2 structure prediction

(Jumper, J. et al. 2021. Nature)

TEMPRO

(Alvarez, J.A.E., Dean, S.N. 2024)

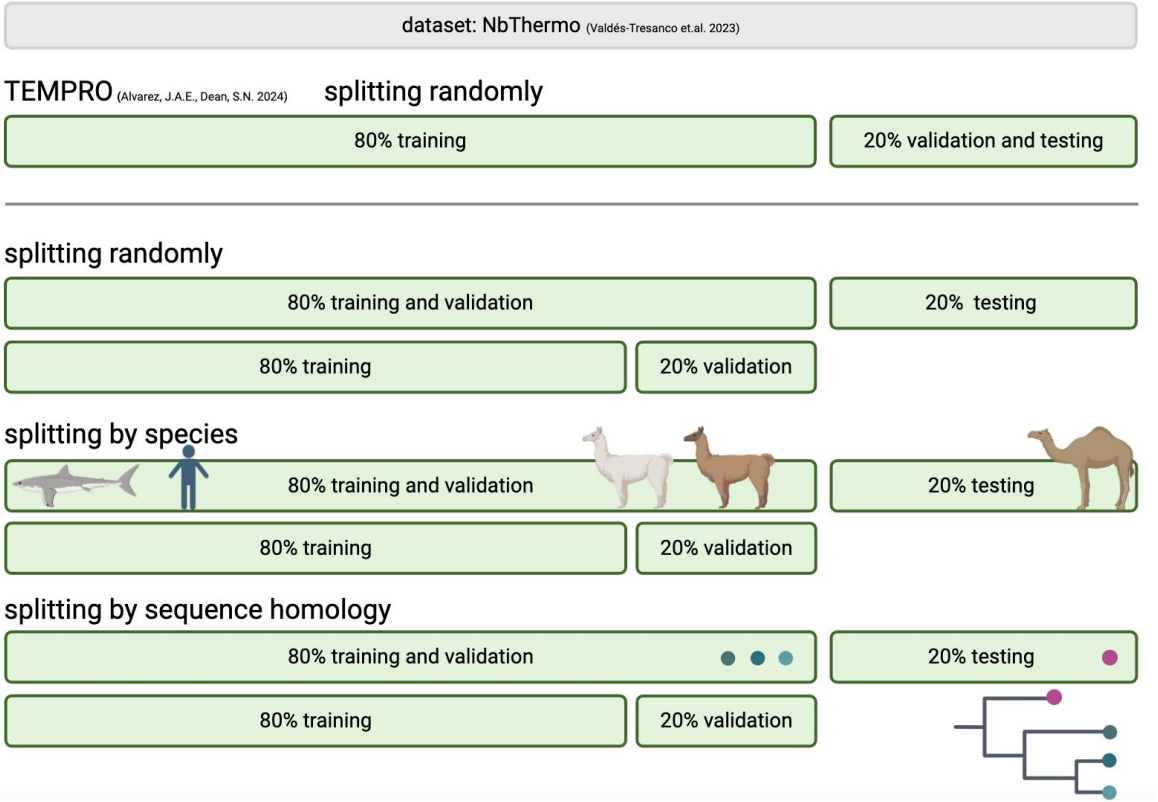
47 ºC

66 ºC

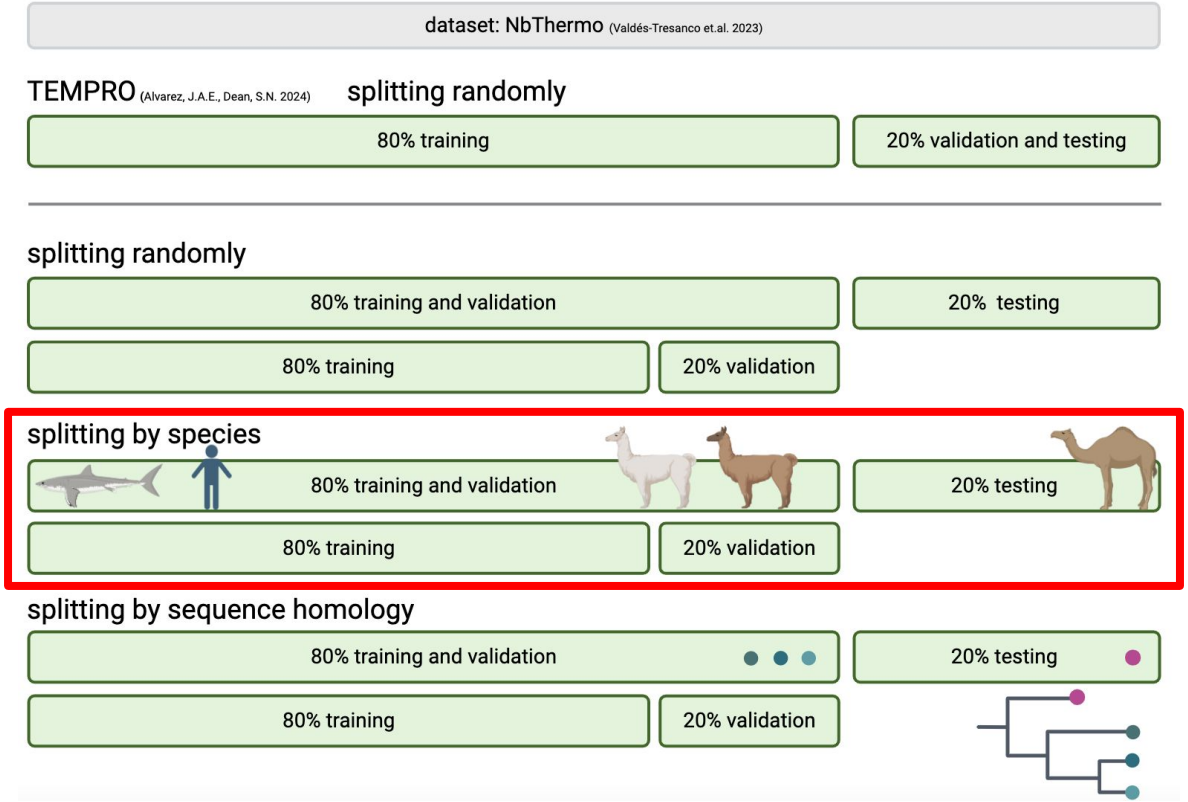
Predicted melting temperature (Tm)

But do we trust this model?

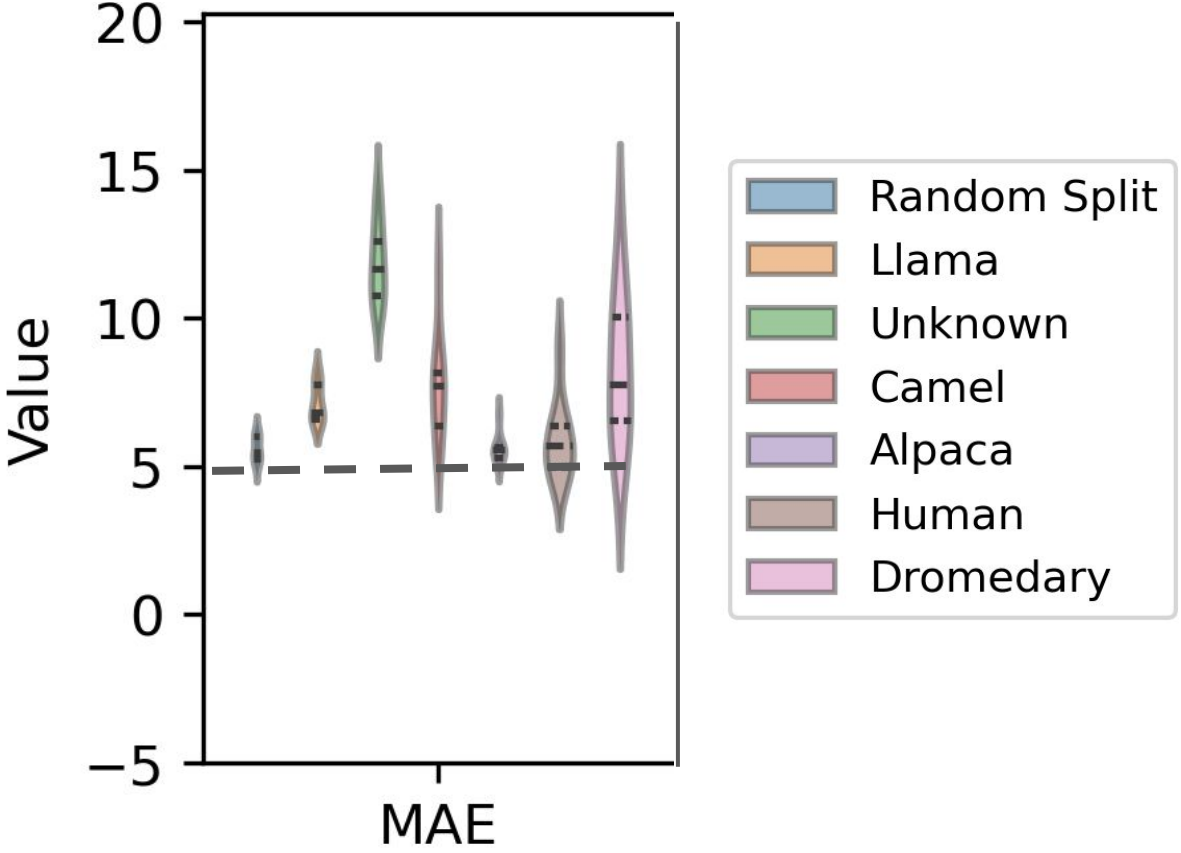
Sensitivity analyses and uncertainty estimates are crucial to trust model predictions.



Sensitivity analyses and uncertainty estimates are crucial to trust model predictions.

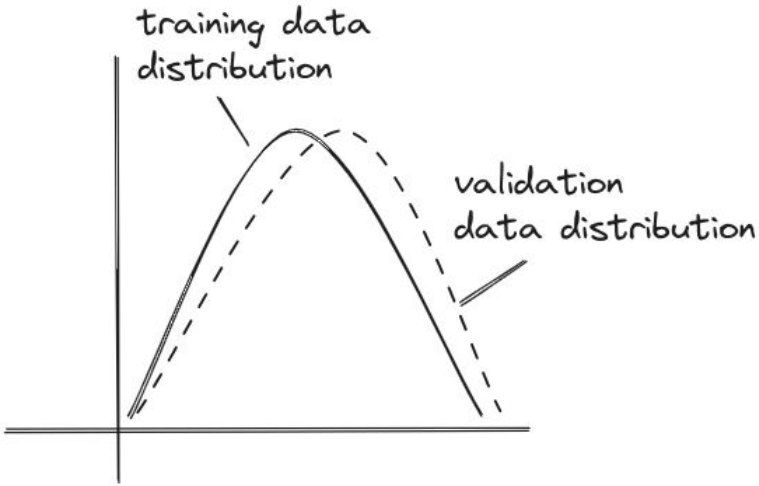


Random split validation underestimates the error for novel species

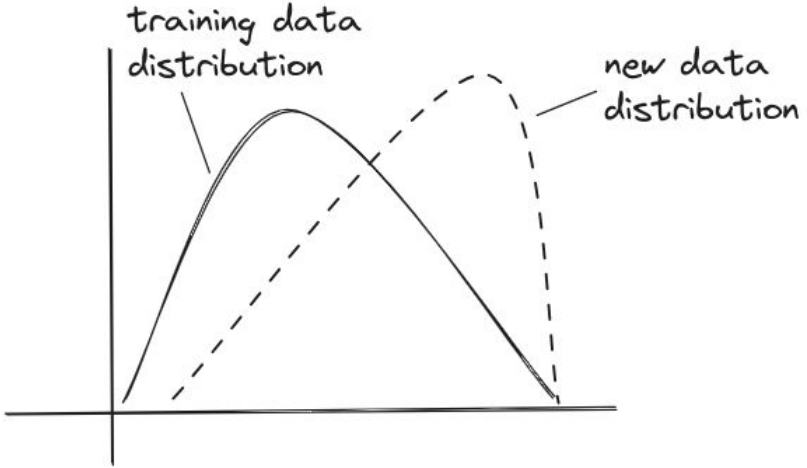


Take-home message

Training: great performance ✓



Inference: wrong predictions ✗



Thank you!

Contributors:
Finnja Becker
Madalina Giurgiu
Dr. Anne Hartebrodt

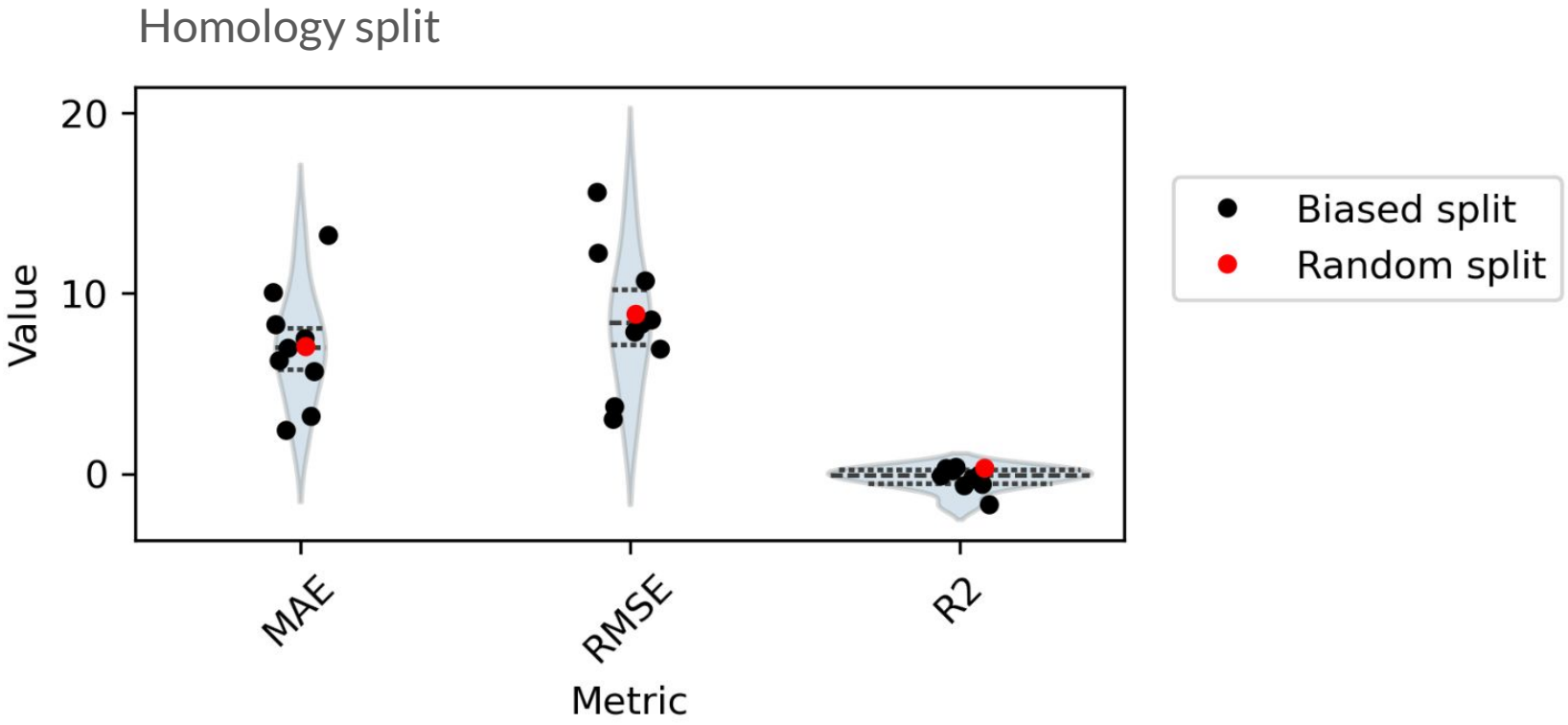


Project advisor: Dr. Ni Fang

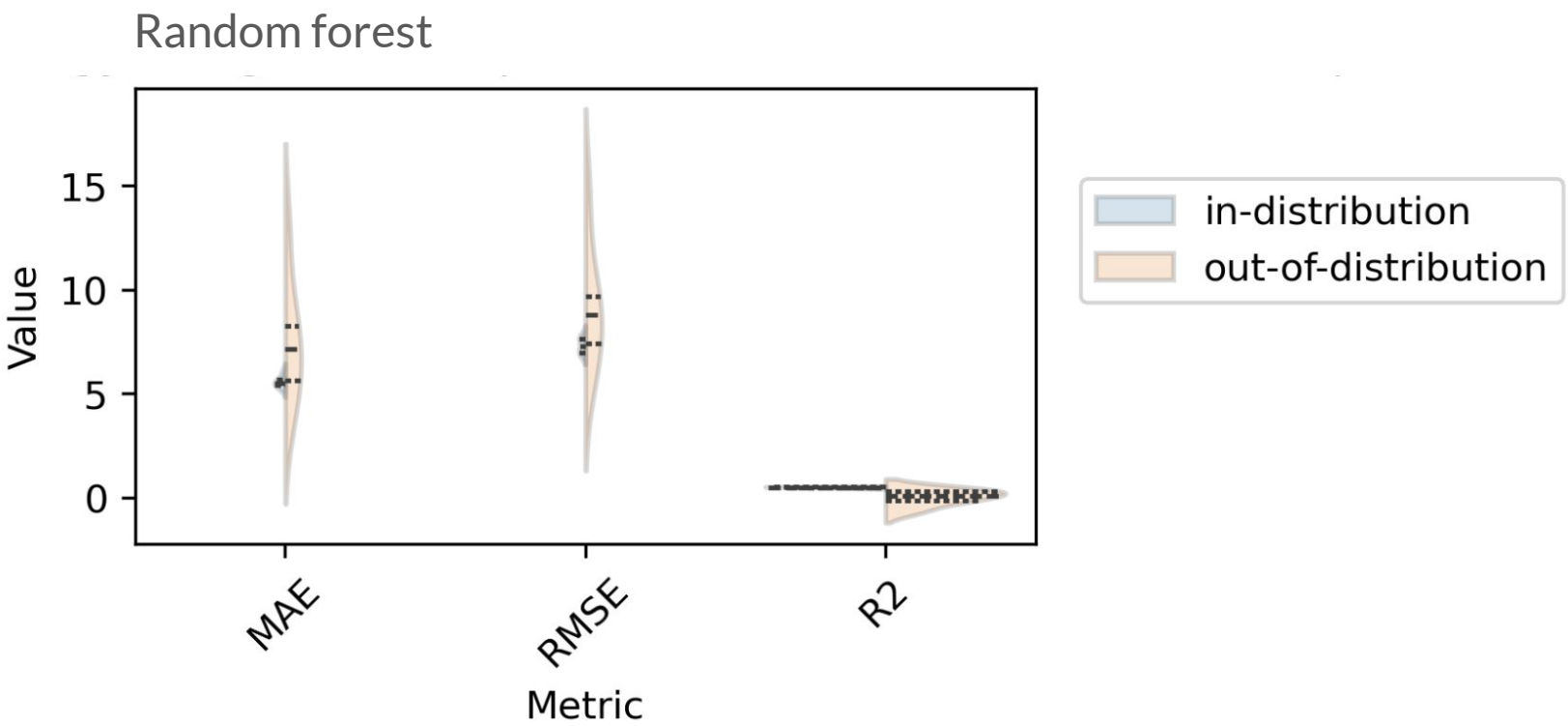
Github repository: https://github.com/AnneHartebrodt/earth-ml-sensitivity

Backup

Validation strategy underestimates the error for novel species and non-homologous sequences



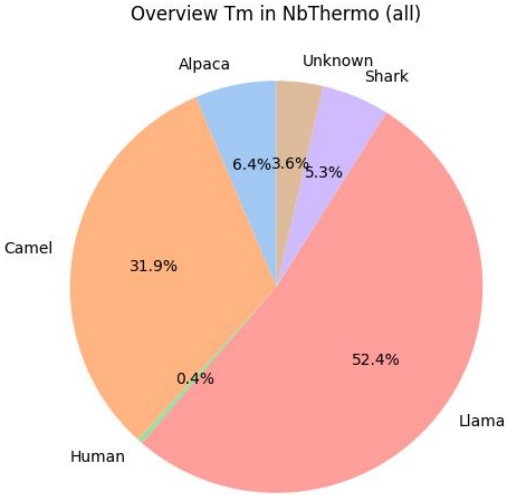
Baseline ML model accuracy

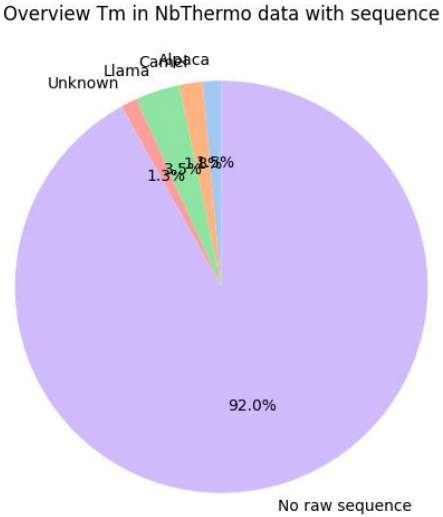


Model improvement ideas

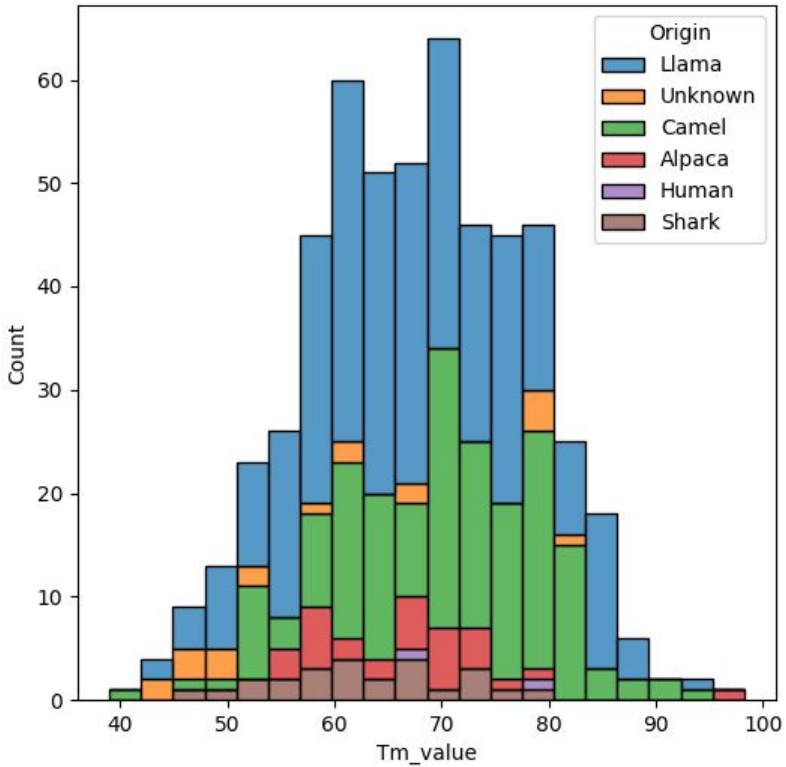
- Ensemble classifier
- More data
- Train different models (e.g. protein melting temp. models) on nanobody data
- Fine-tuning (use protein data to train larger model, fine-tune this model using the limited nanobody data)

NbThermo


Overview Tm in NbThermo (all)
Alpaca
Unknown
Shark
6.4%
3.6%
5.3%
Camel
31.9%
52.4%
0.4%
Llama
Human


Overview Tm in NbThermo data with sequence
Unknown
Llama
Camel
Alpaca
1.3%
92.0%
No raw sequence

NbThermo:


Origin
Llama
Unknown
Camel
Alpaca
Human
Shark
Count
0
10
20
30
40
50
60
40
50
60
70
80
90
100
Tm_value