TS037-Wallner group
AlphaFold with aggressive sampling

Results from the Protein Assembly Category

Björn Wallner
Linköping University

LiU
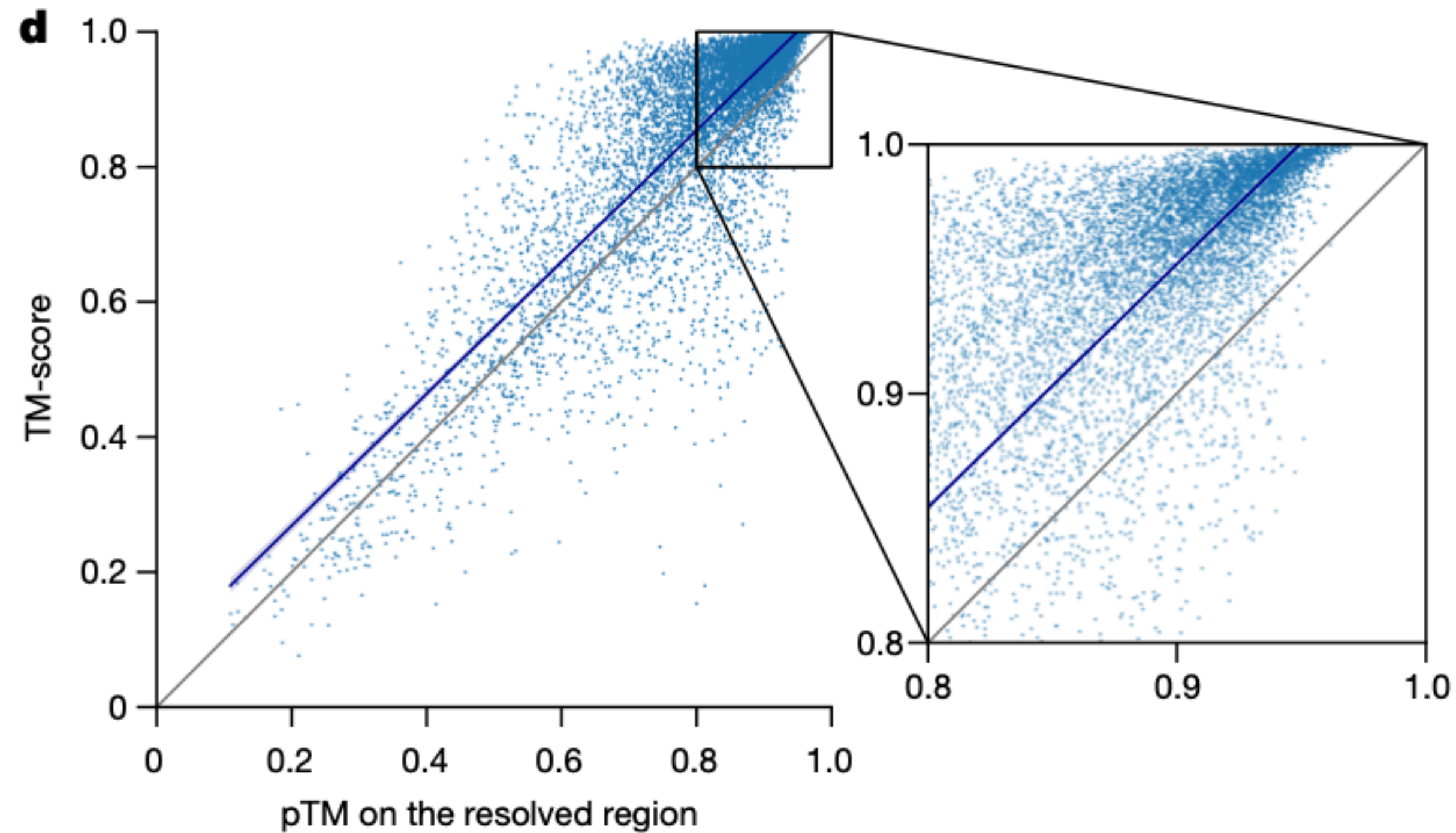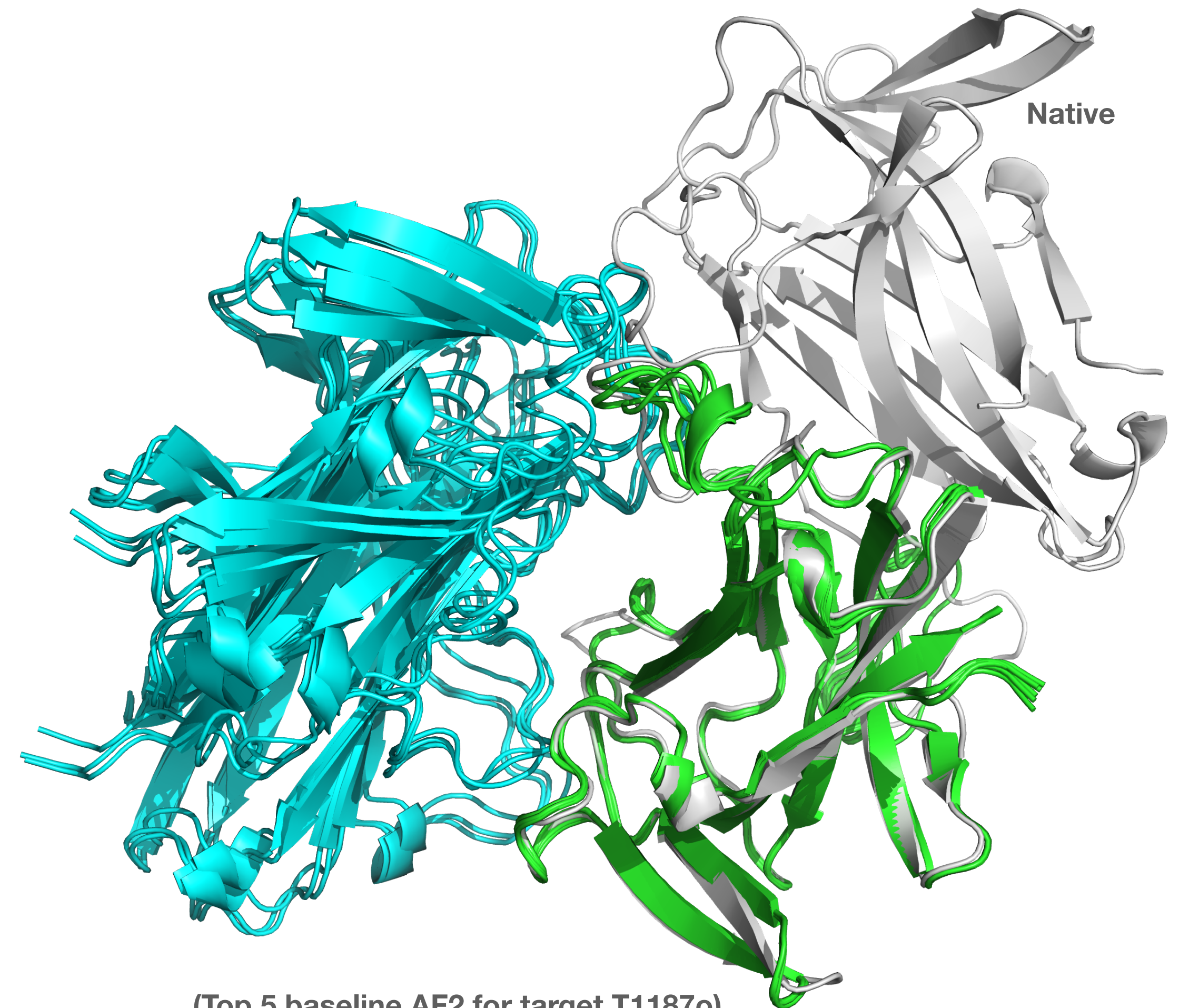
LINKÖPING UNIVERSITY

Motivation for the method

The scoring function in AlphaFold is stellar
In particular for high quality models

Model diversity is NOT stellar
In particular, when it comes to sampling different interfaces


(Jumper, et al, 2021)

Native

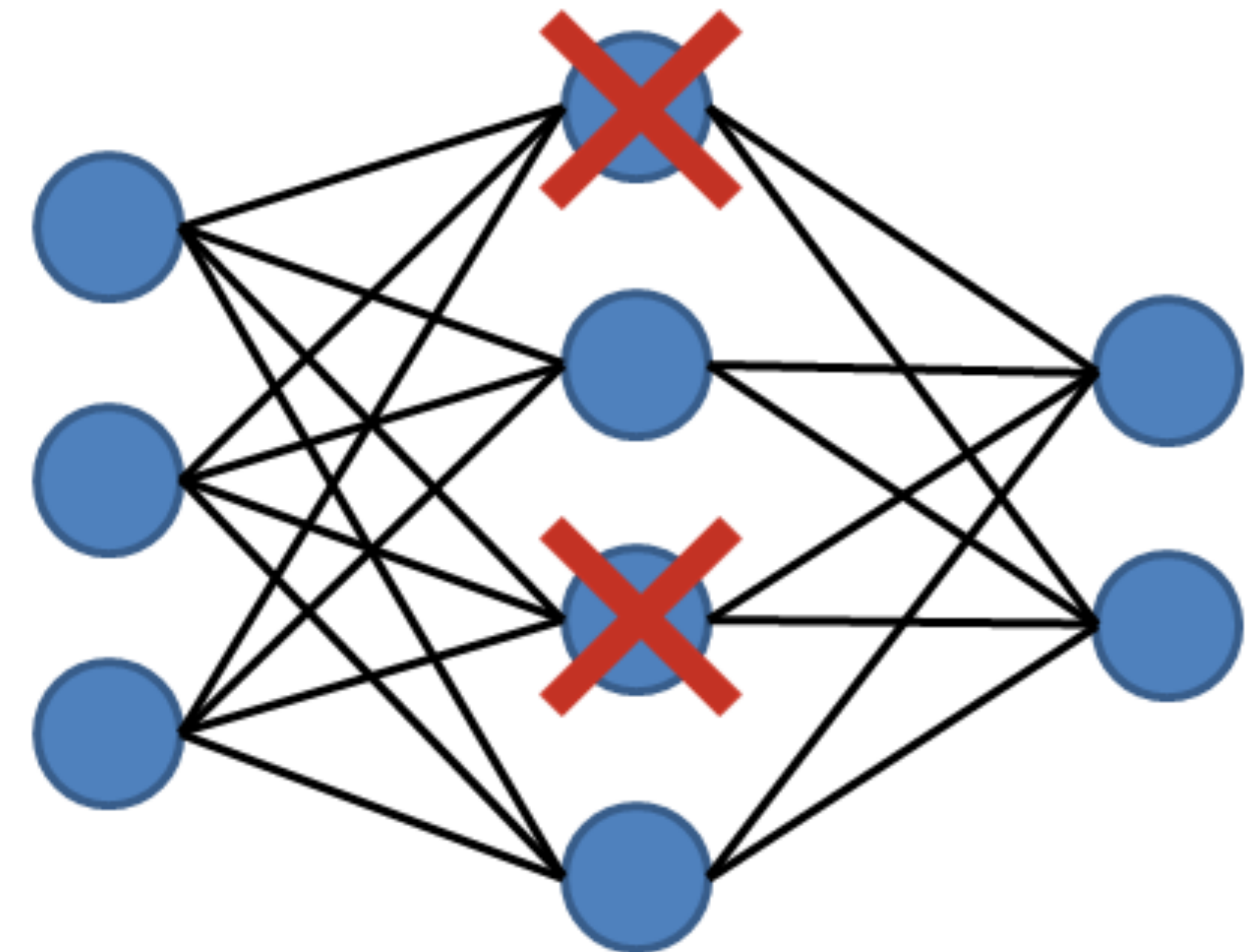(Top 5 baseline AF2 for target T1187o)

Ways to achieve more model diversity in AF2

- Generate more structures with a different random seed
  - *multimer_v2* default generates 25 models compared to 5 for *multimer_v1*
- Change the number of recycles
- Perturb the input and sample more structures
  - Sub-sample the MSA
  - Change the templates
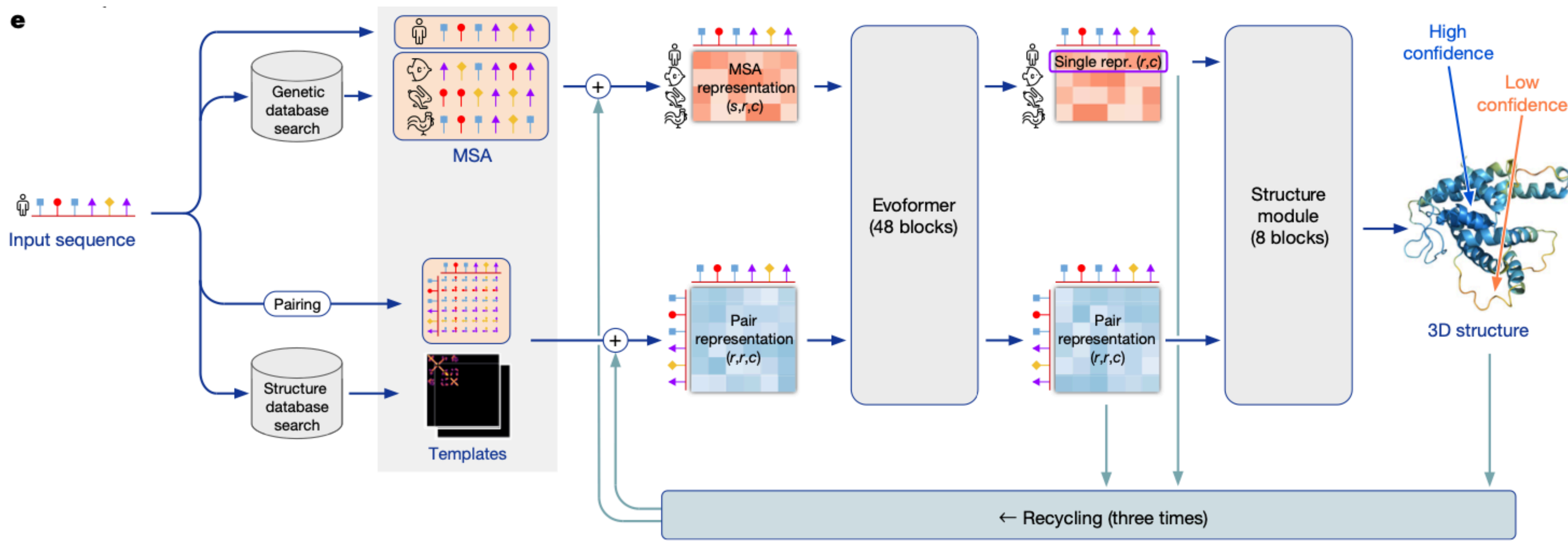 - **Turn on the dropout layers at inference**

(Jumper, et al, 2021)
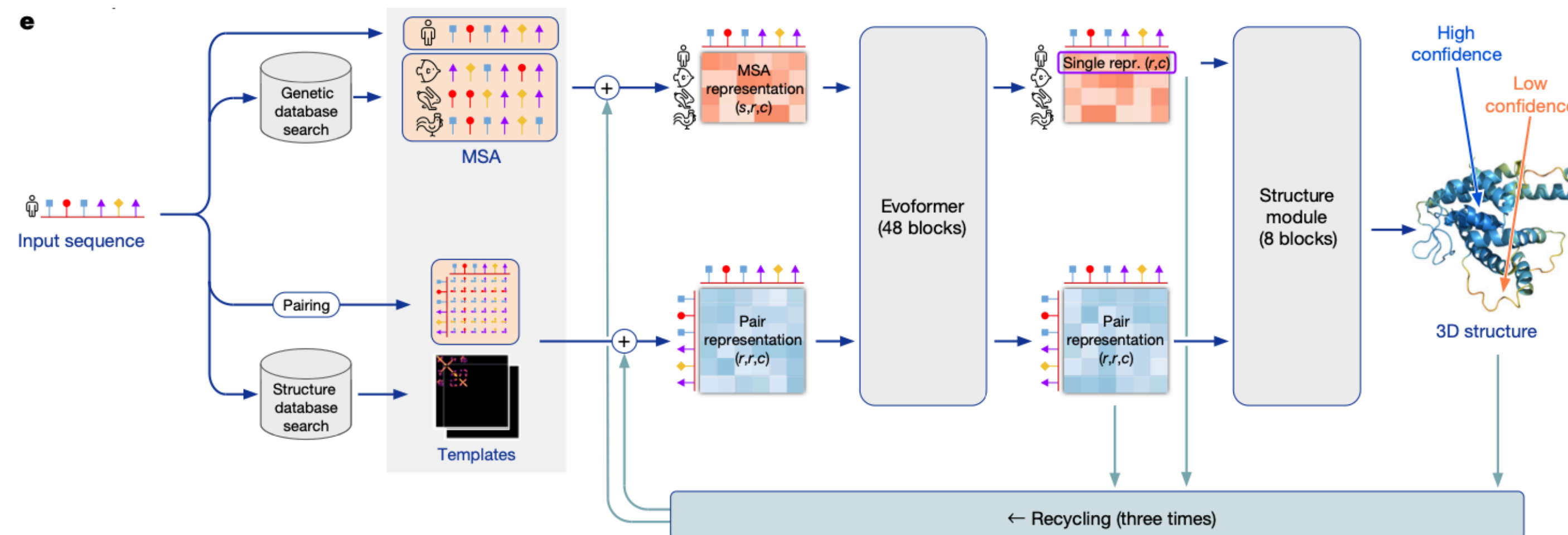
Dropout layers

- Dropout is a regularisation technique used during training normally.

- Randomly drop "dropout_rate" number of neurons, at each iteration in the training.

- Improves generalisation

  - Effectively trained on smaller networks

  - Redundancy

- Typically, at inference, the dropout_rate is set to zero and the full network is used.

- **By allowing dropout at inference we can sample the uncertainties in the input data**
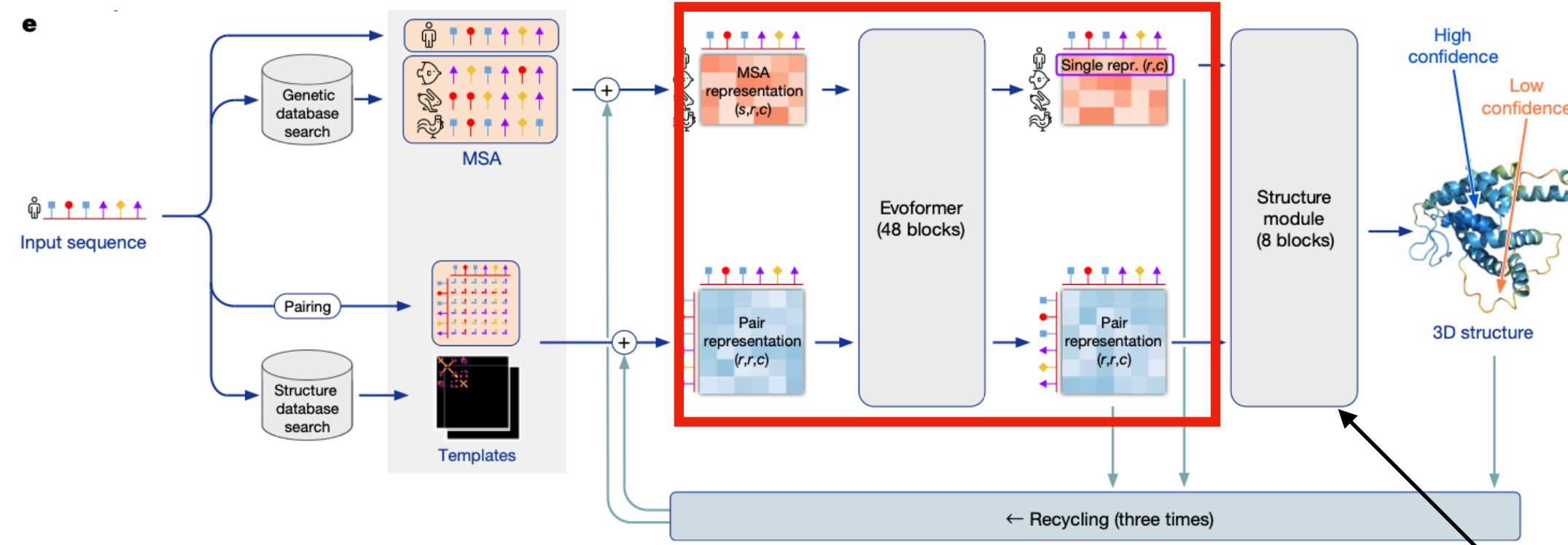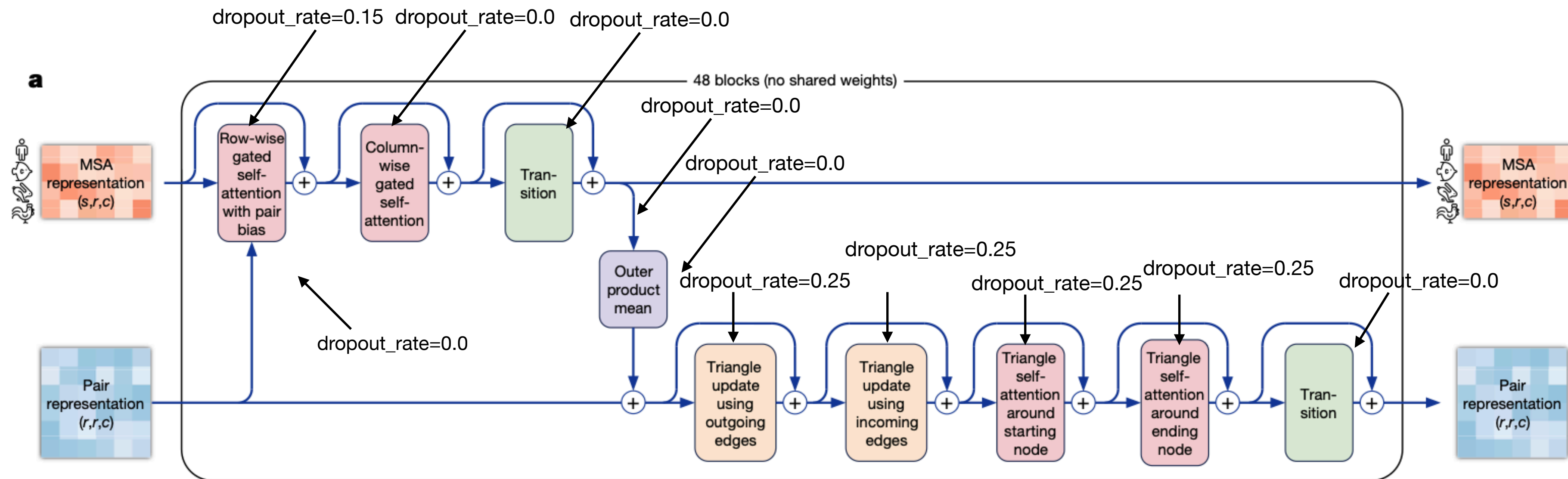

Network with dropout_rate=0.5

Dropout in the AlphaFold Network
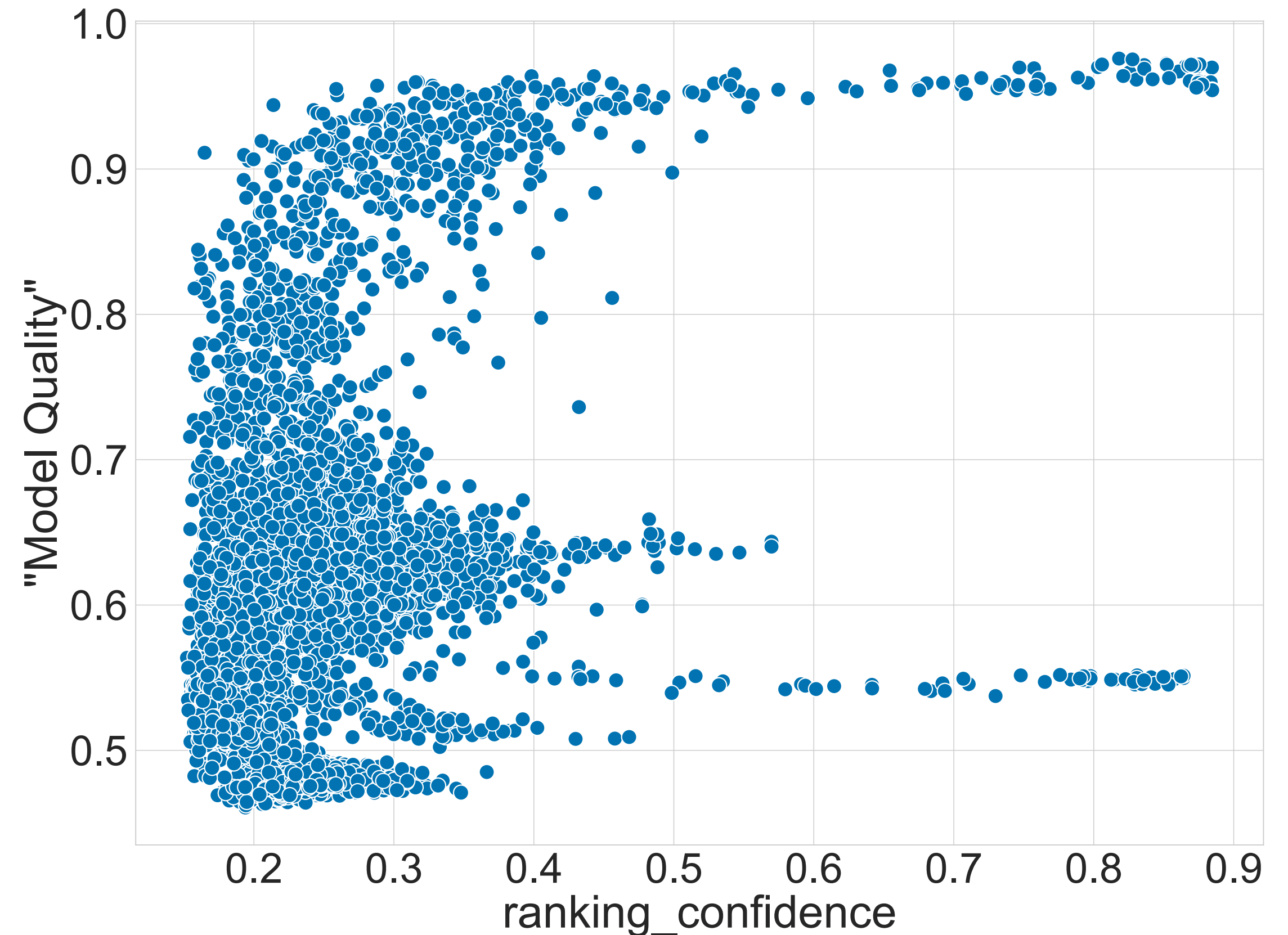

(Jumper et al Nature 2021)

e

Input sequence

Genetic database search

Pairing

Structure database search

MSA

Templates

MSA representation (s,r,c)

Pair representation (r,r,c)

Evoformer (48 blocks)

Single repr. (r,c)

Pair representation (r,r,c)

Structure module (8 blocks)

High confidence

Low confidence

3D structure

← Recycling (three times)

dropout_rate=0.1

a

48 blocks (no shared weights)

dropout_rate=0.15   dropout_rate=0.0   dropout_rate=0.0

dropout_rate=0.0

MSA representation (s,r,c)

Row-wise gated self-attention with pair bias

Column-wise gated self-attention

Tran-sition

MSA representation (s,r,c)

Outer product mean

dropout_rate=0.0

dropout_rate=0.0

dropout_rate=0.25

dropout_rate=0.25

dropout_rate=0.25

dropout_rate=0.25

dropout_rate=0.0

Pair representation (r,r,c)

Triangle update using outgoing edges

Triangle update using incoming edges

Triangle self-attention around starting node

Triangle self-attention around ending node

Tran-sition

Pair representation (r,r,c)

(Jumper et al Nature 2021)

Method Overview

- Use AlphaFold to generate models (full_dbs)

- Turn on dropout at inference

    1. All dropout_rates at default

    2. No dropout in the structural module

- Generate many structures with a different random seed

- Use both:

    - *multimer_v1*

    - *multimer_v2*

- Run with and without templates

- Run with an increased number of recycles

- Select based on the ranking_score

Settings used and motivation

Model	Templates	Dropout	Recycles	nstruct
multimer_v1	Yes	Yes	3	5x200
multimer_v1	No	Yes*	3	5x200
multimer_v1	No	Yes*	21	5x200
multimer_v2	Yes	Yes	3	5x200
multimer_v2	No	Yes*	3	5x200
multimer_v2	No	Yes*	9	5x200

* but not in structural module

Improving peptide-protein docking with AlphaFold-Multimer using forced sampling

Isak Johansson-Åkhe and Björn Wallner*

Division of Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden

Larger spread in model quality when using dropout

multimer_v1 "likes" more recycles than v2

multimer_v1 produces models with higher DockQ

Amount of sampling per target

~18 GPUs during CASP
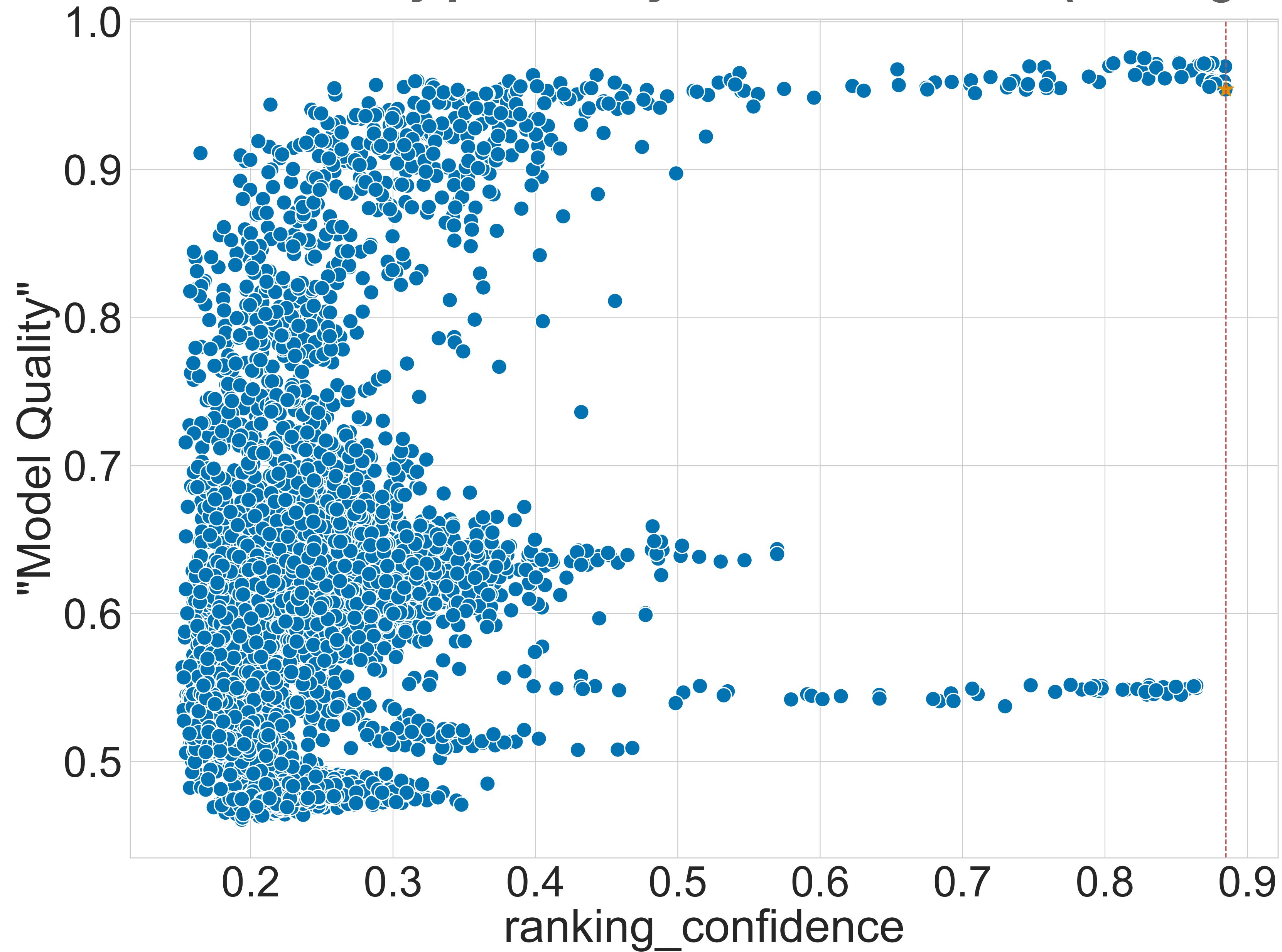Cost ~CASP15 conference fee
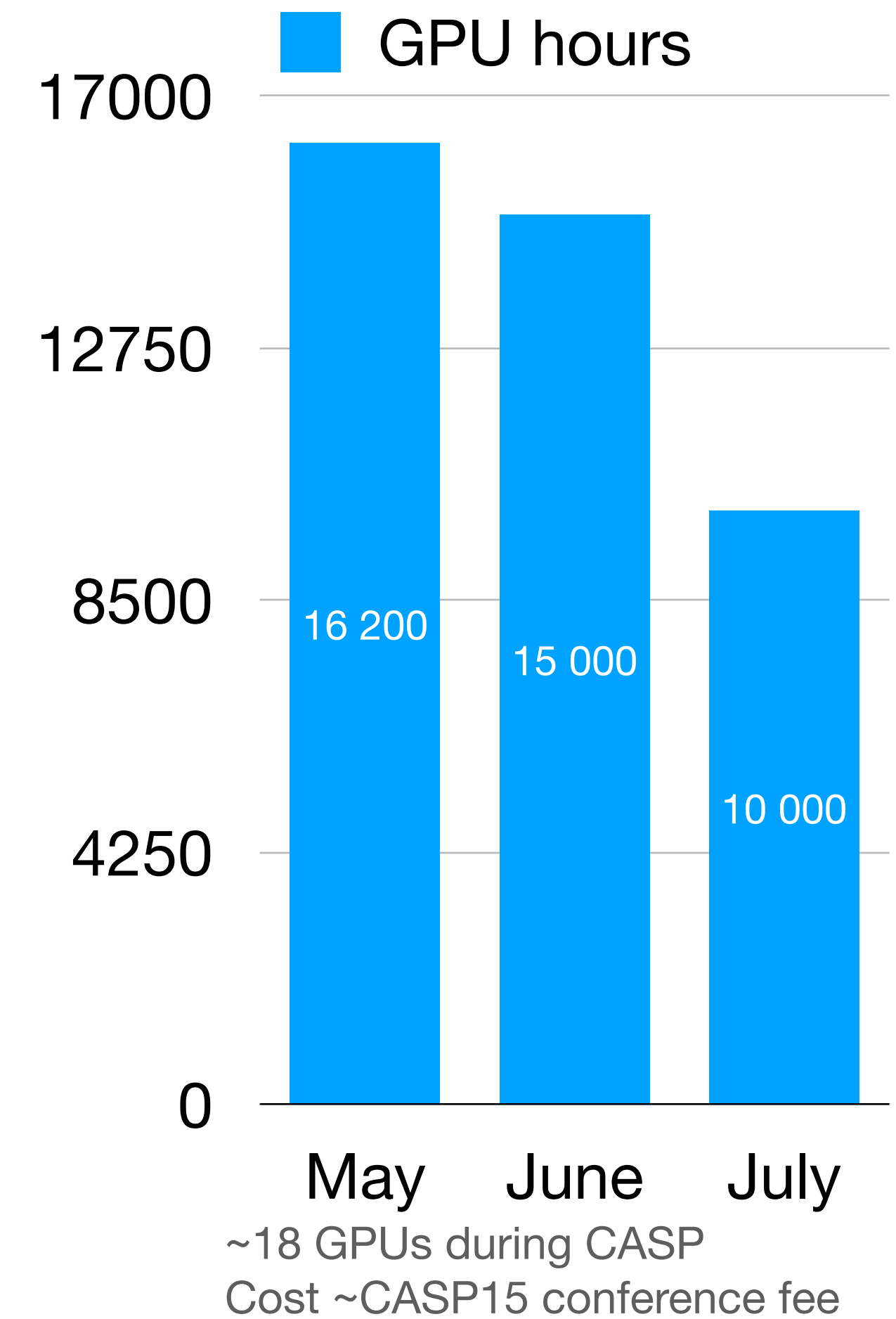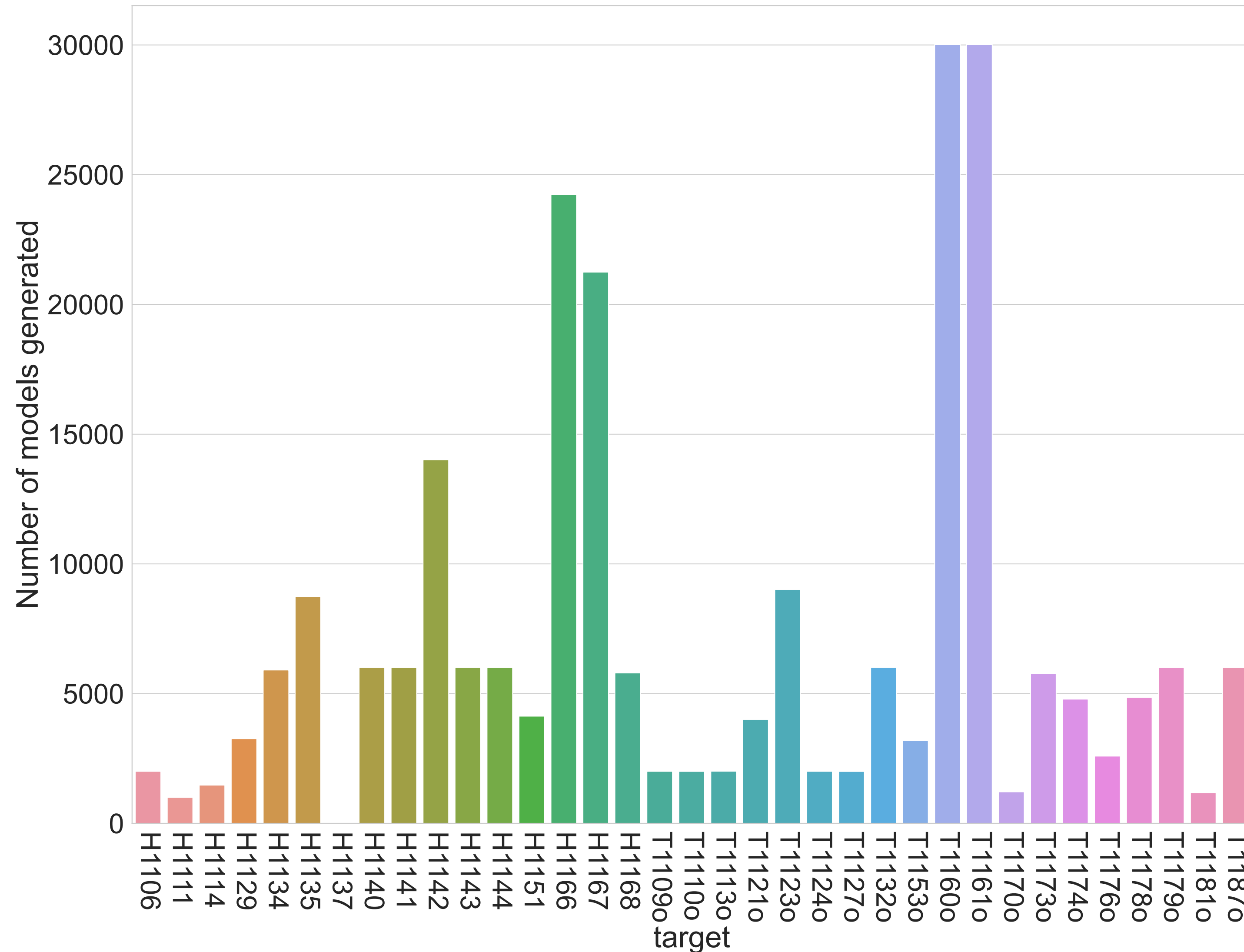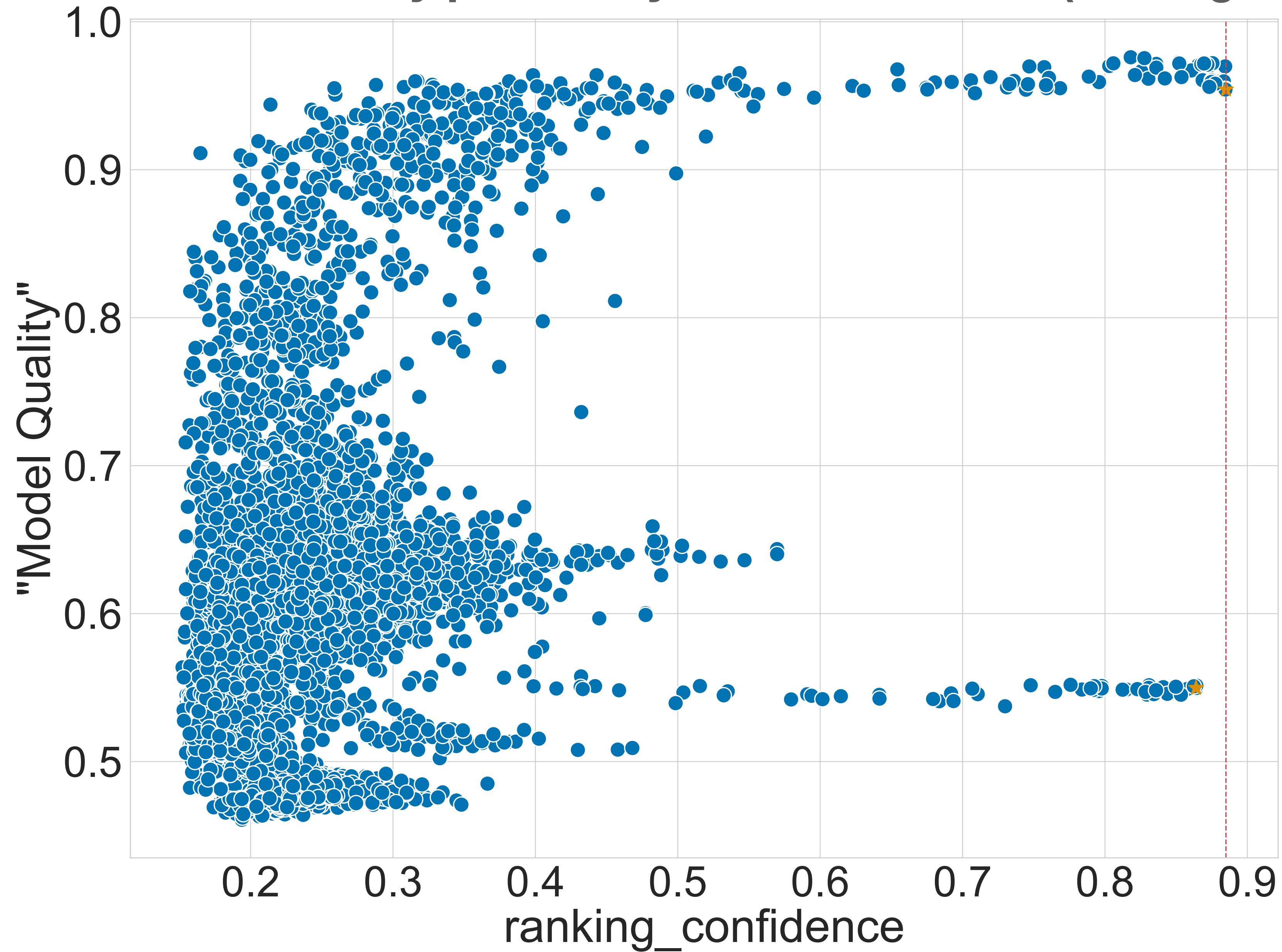
Selection of models for submission

1
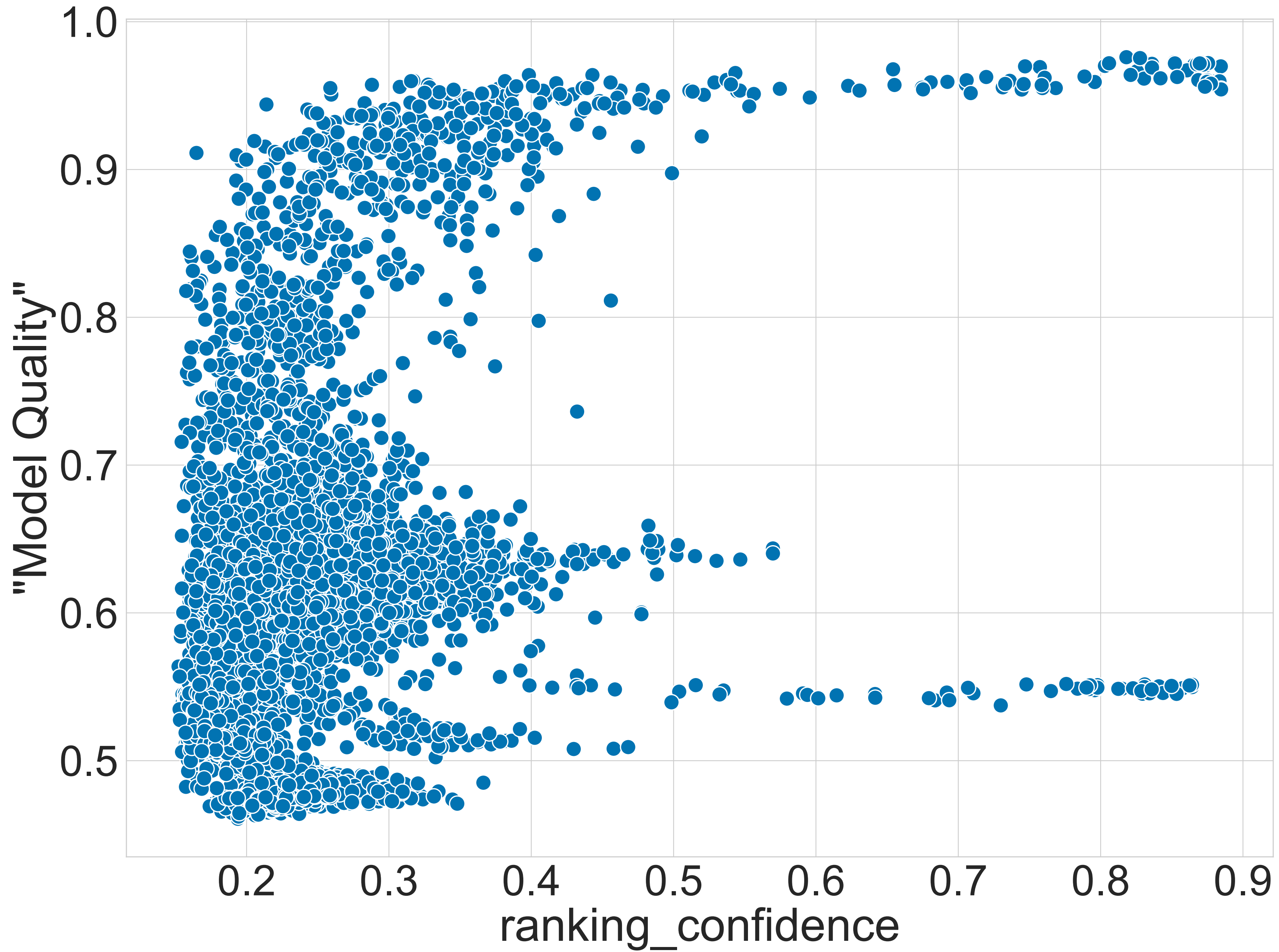
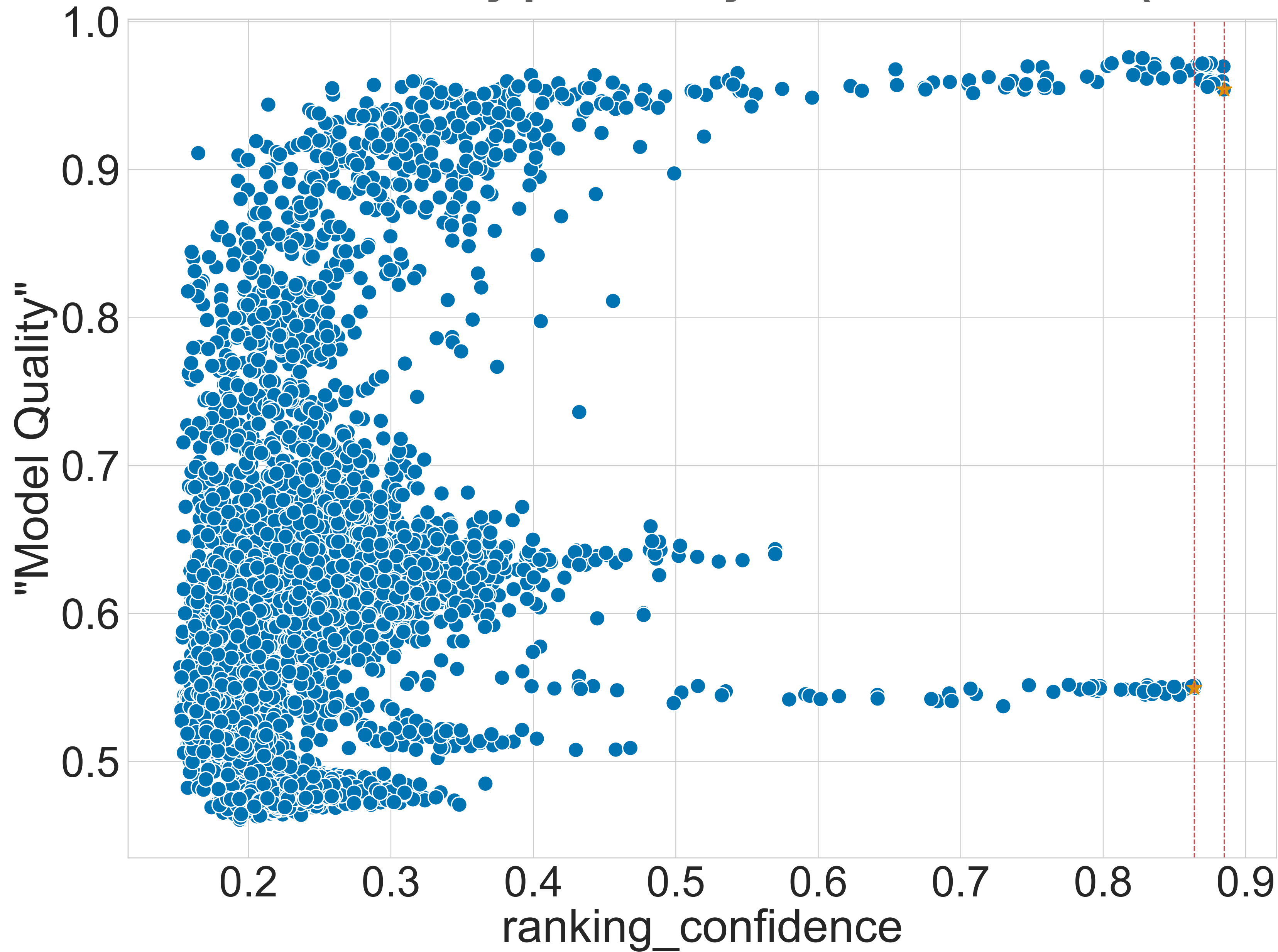# 1. Select the model with the highest score from the pool
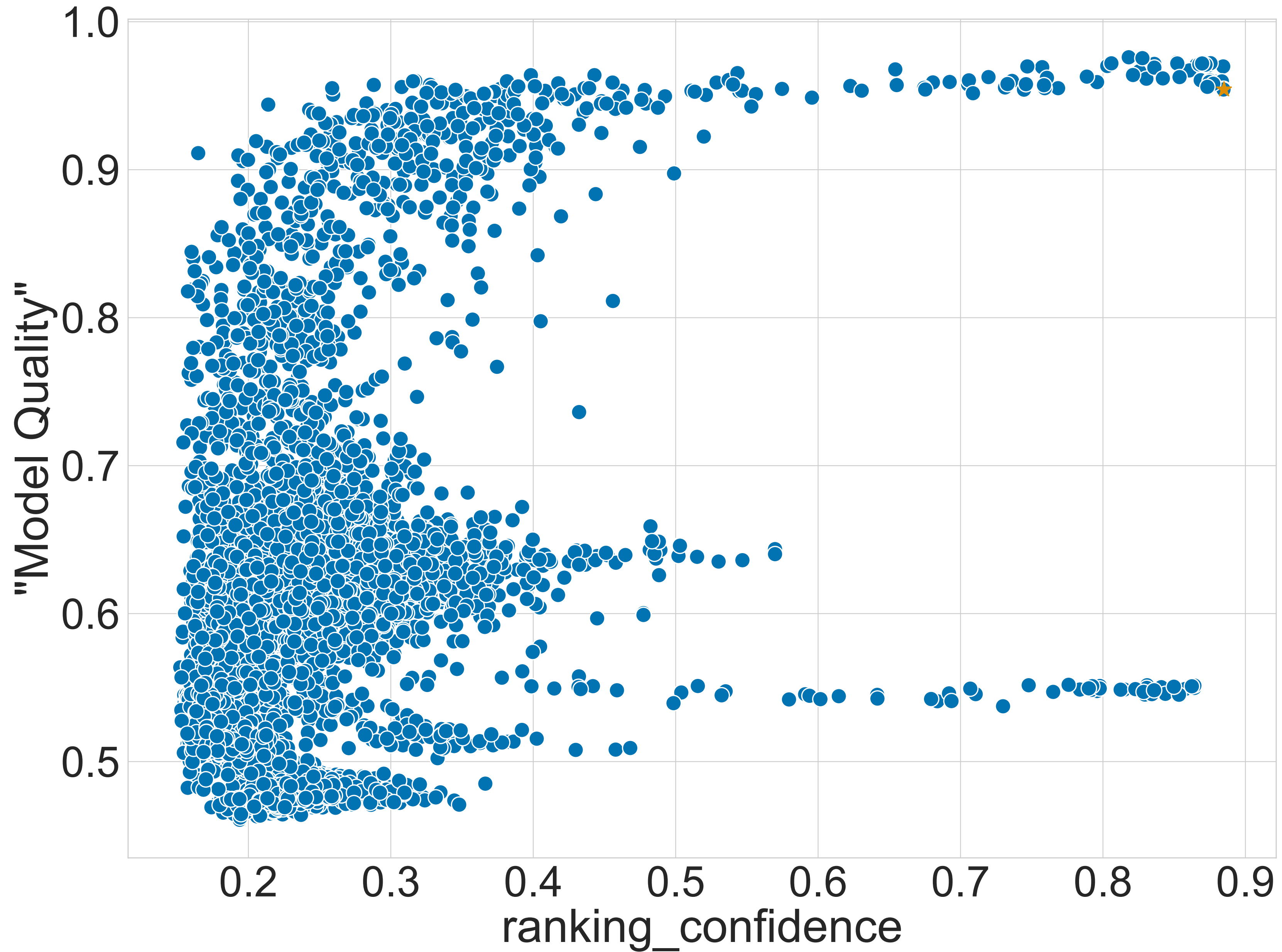
2. Select the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)
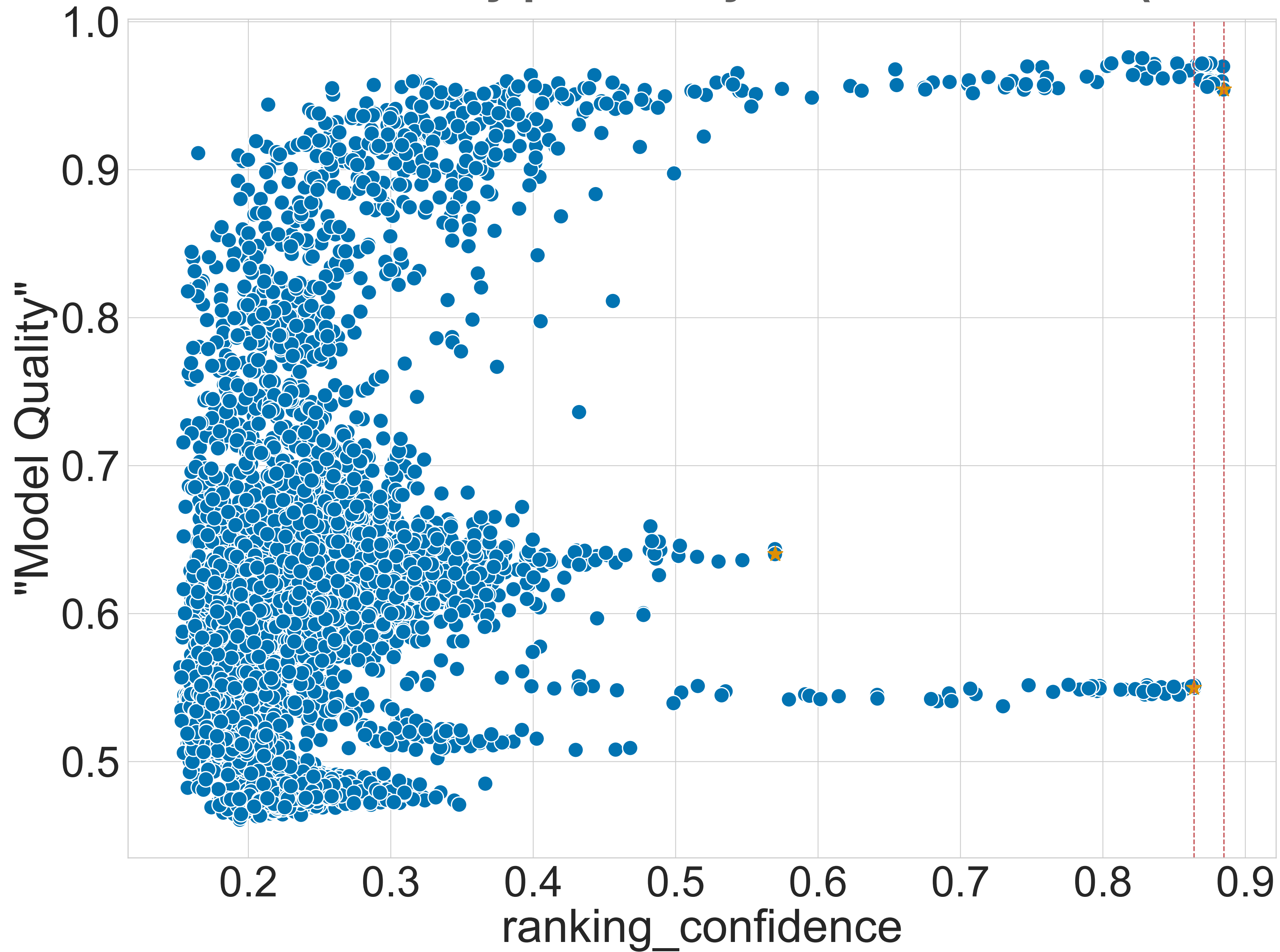
2. Select the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)
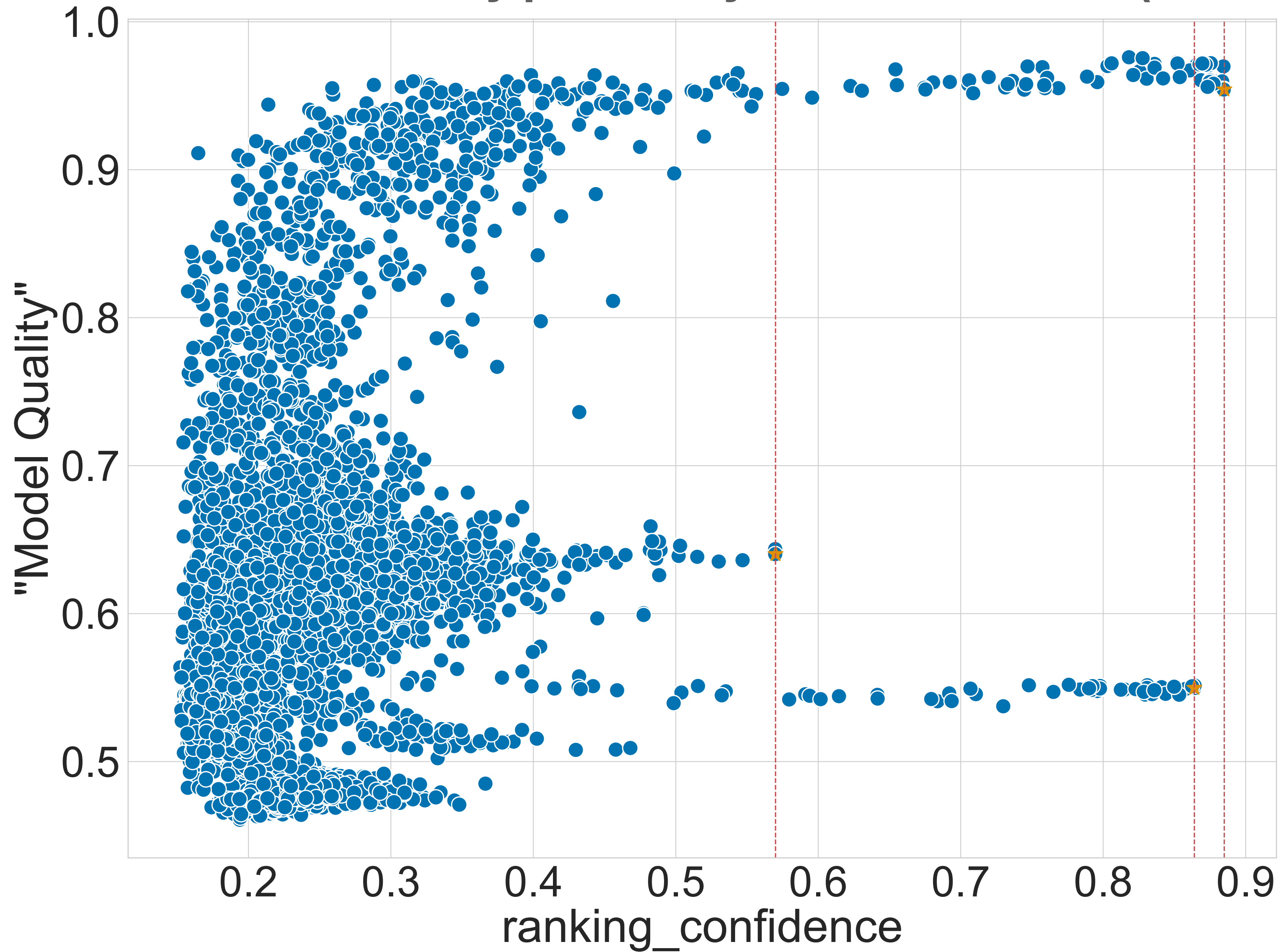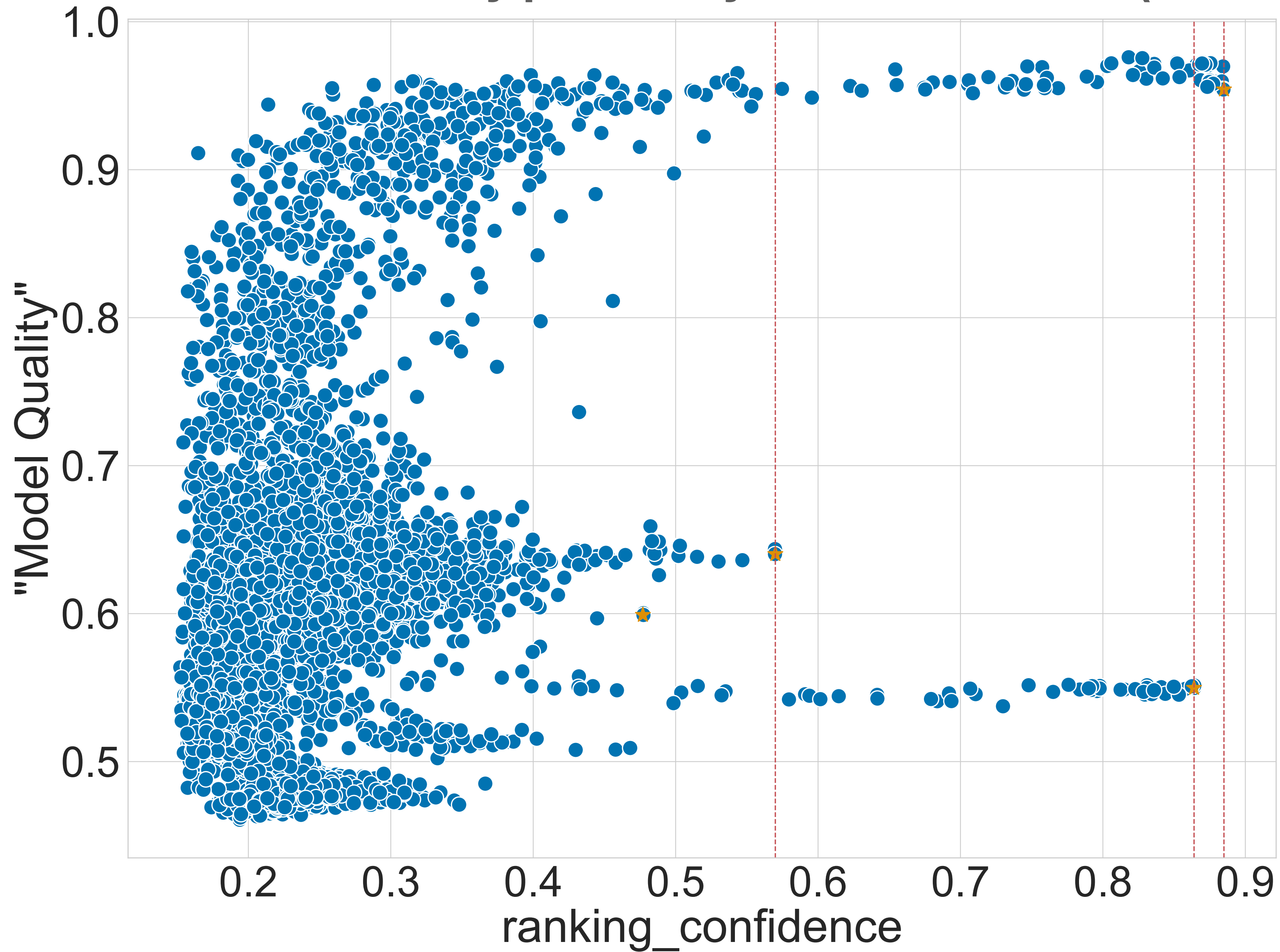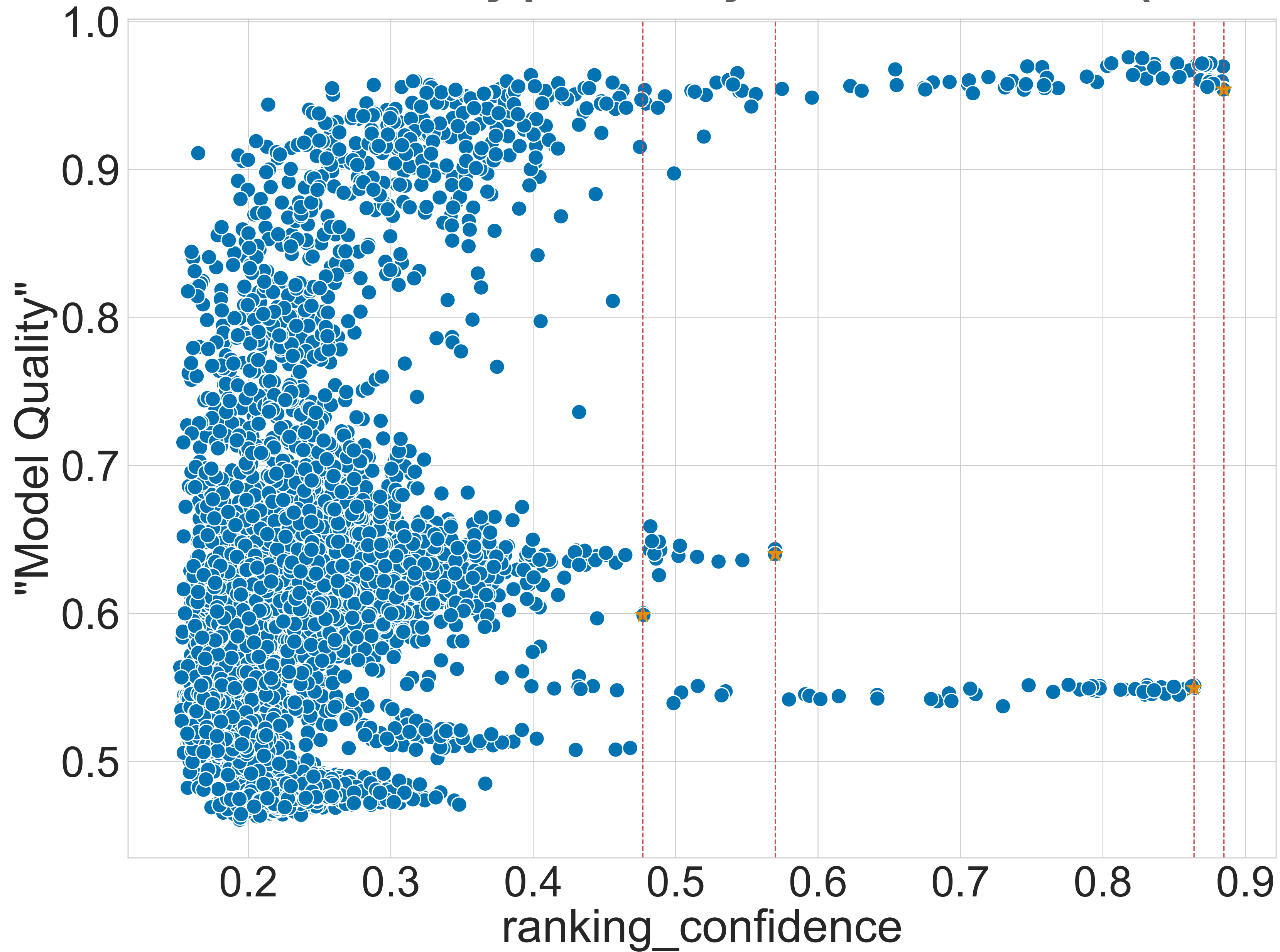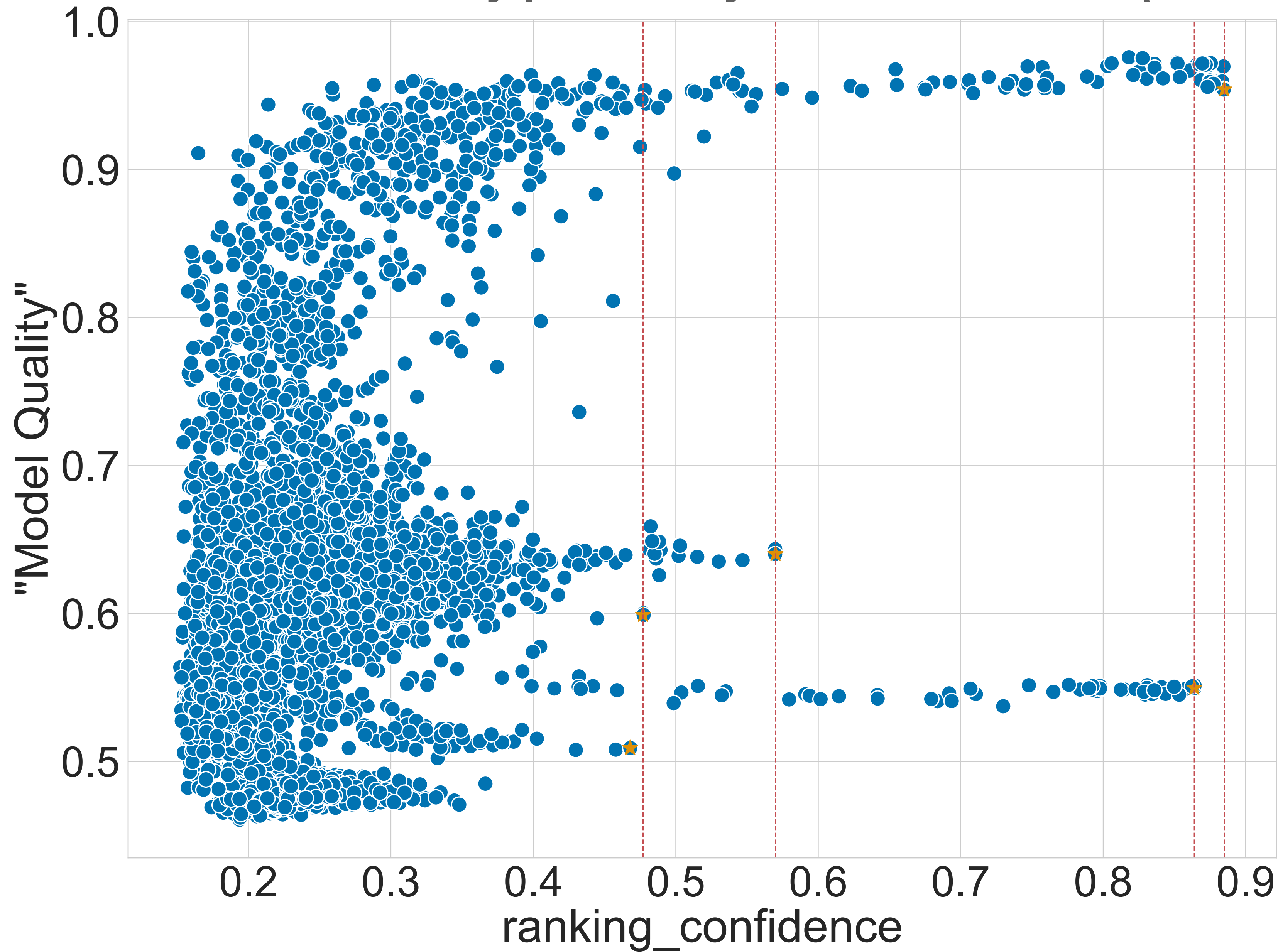
3. Continue with selecting the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)

3. Continue with selecting the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)

3. Continue with selecting the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)

3. Continue with selecting the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)

3. Continue with selecting the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)

3. Continue with selecting the model with the highest score from the pool, NOT similar to any previously selected models (MMalign<0.8)

4. Until five models are selected, or the pool is exhausted
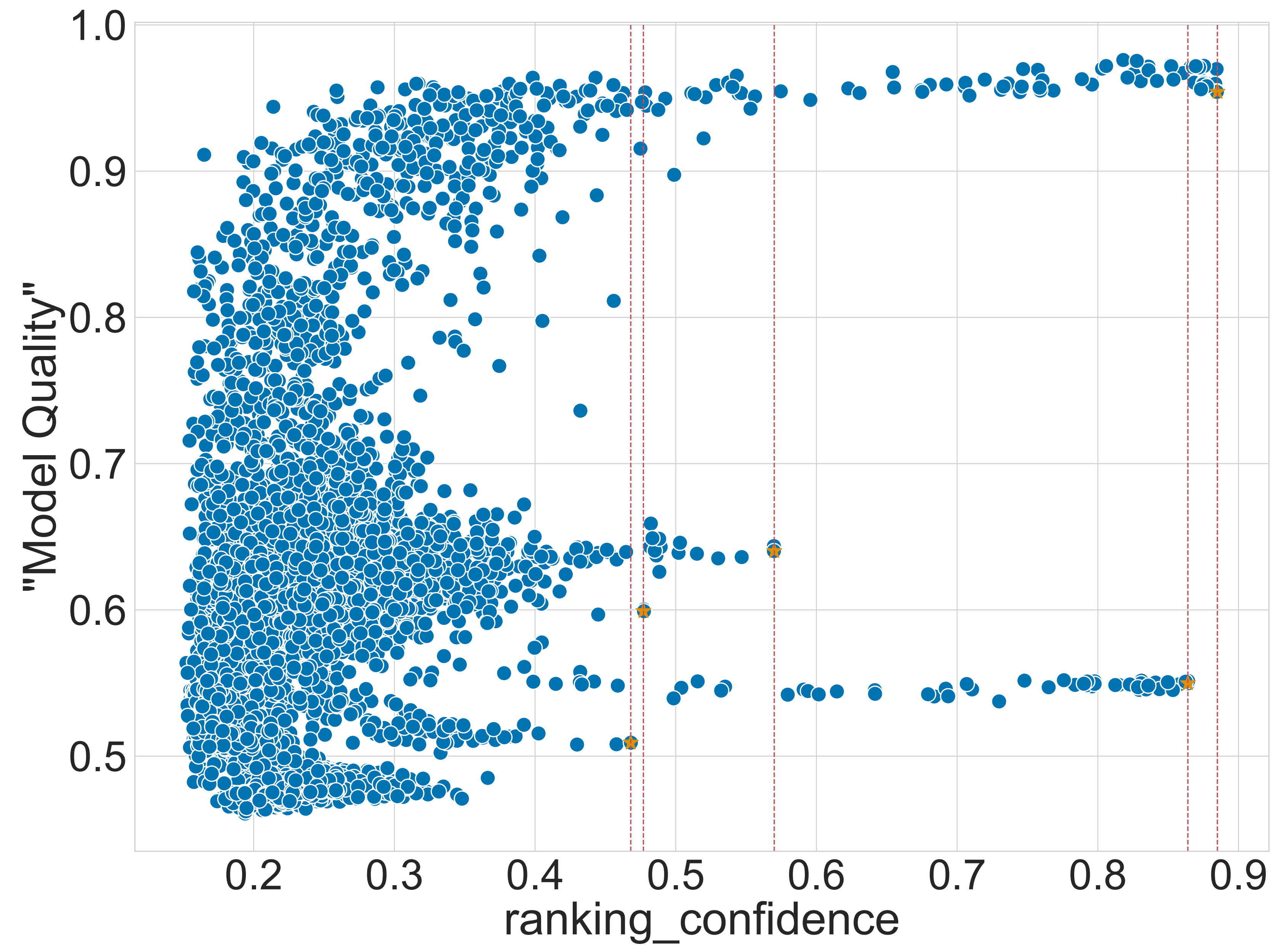
Origin of rank 1 models

Model_v	Templates	Dropout	Recycles	nstruct
multimer_v1	Yes	Yes	3	5x200
multimer_v1	No	Yes*	3	5x200
multimer_v1	No	Yes*	21	5x200
multimer_v2	Yes	Yes	3	5x200
multimer_v2	No	Yes*	3	5x200
multimer_v2	No	Yes*	9	5x200

- 50% from v1 and 50% from v2

- 50% used templates

- 30% used no templates and recycles

- 20% used no templates
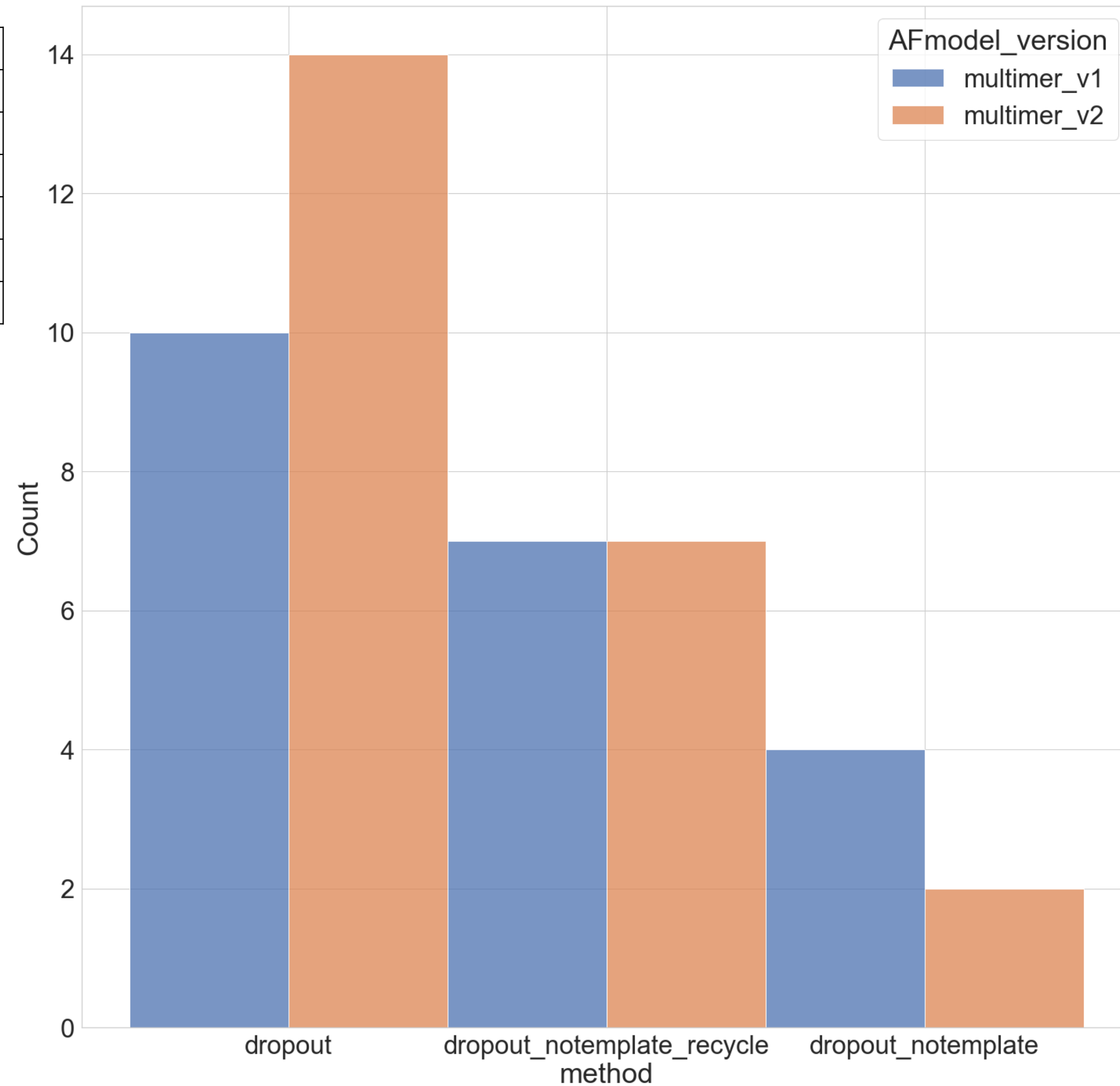
Comparison between the rank 1 performance to individual versions and combinations
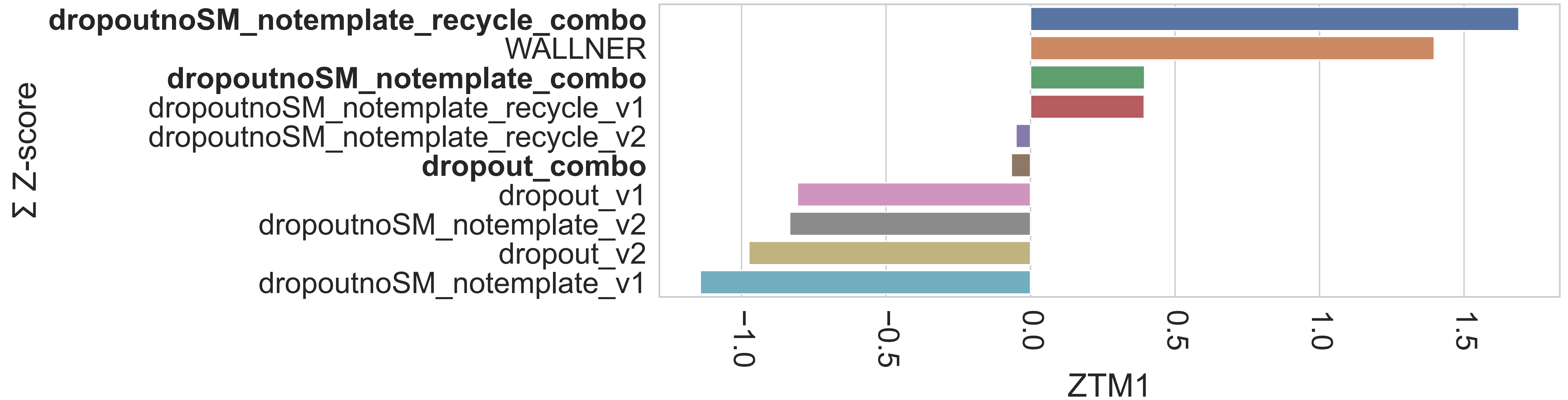
Model_v	Templates	Dropout	Recycles	nstruct	Name
multimer_v1	Yes	Yes	3	5x200	dropout_v1
multimer_v1	No	Yes*	3	5x200	dropoutnoSM_notemplate_v1
multimer_v1	No	Yes*	21	5x200	dropoutnoSM_notemplate_recycle_v1
multimer_v2	Yes	Yes	3	5x200	dropout_v2
multimer_v2	No	Yes*	3	5x200	dropoutnoSM_notemplate_v2
multimer_v2	No	Yes*	9	5x200	dropoutnoSM_notemplate_recycle_v2

Performance overall

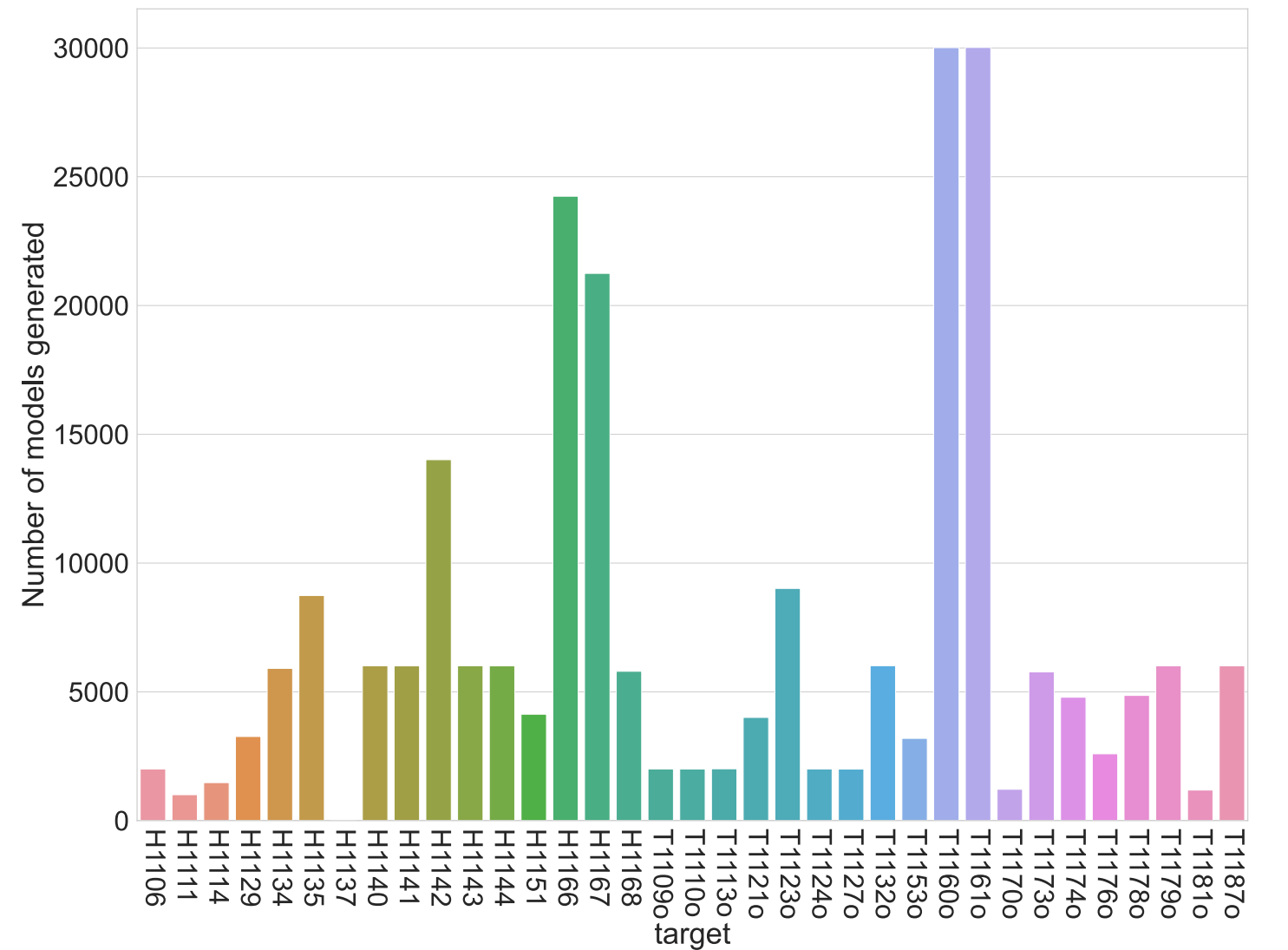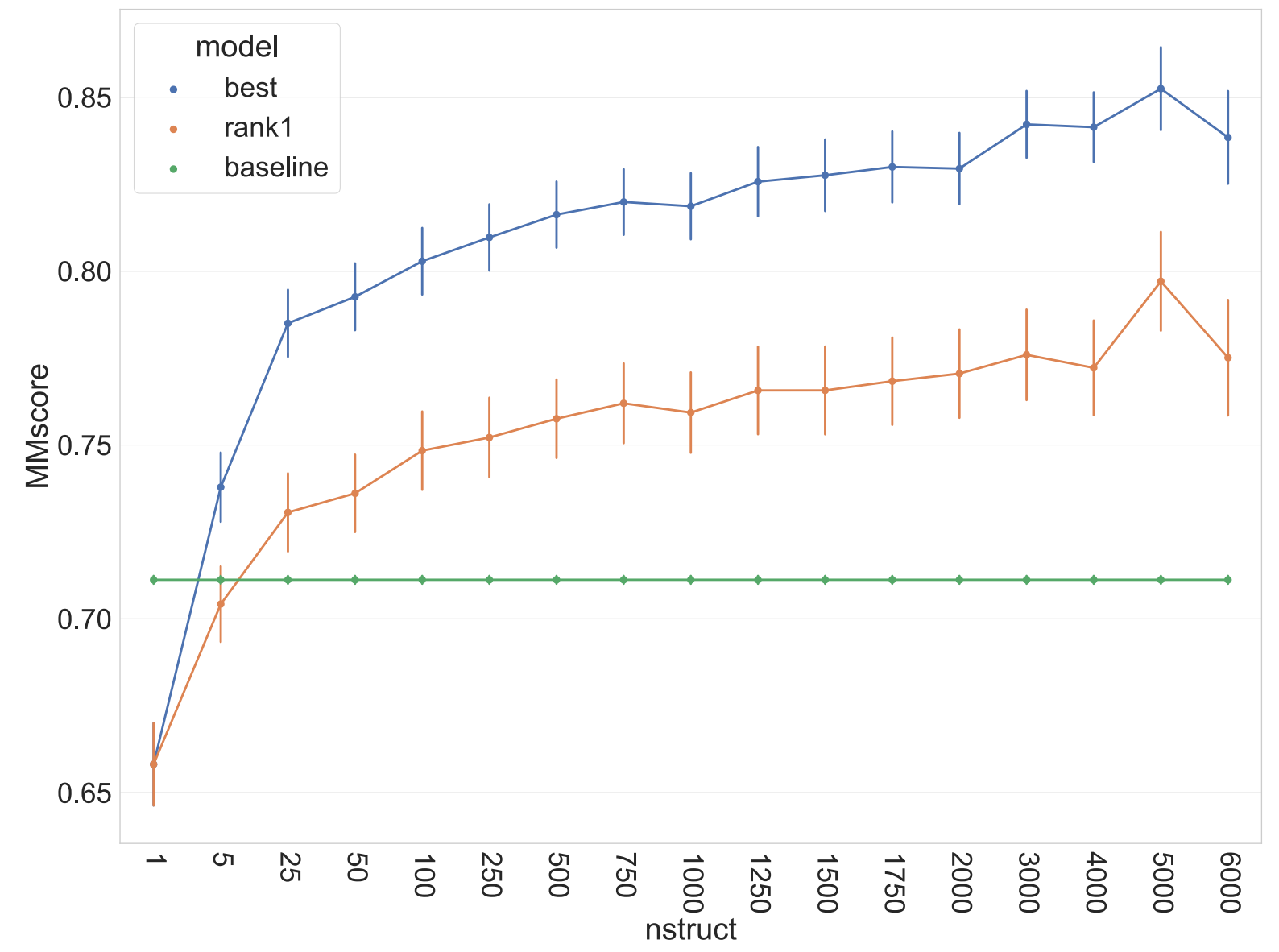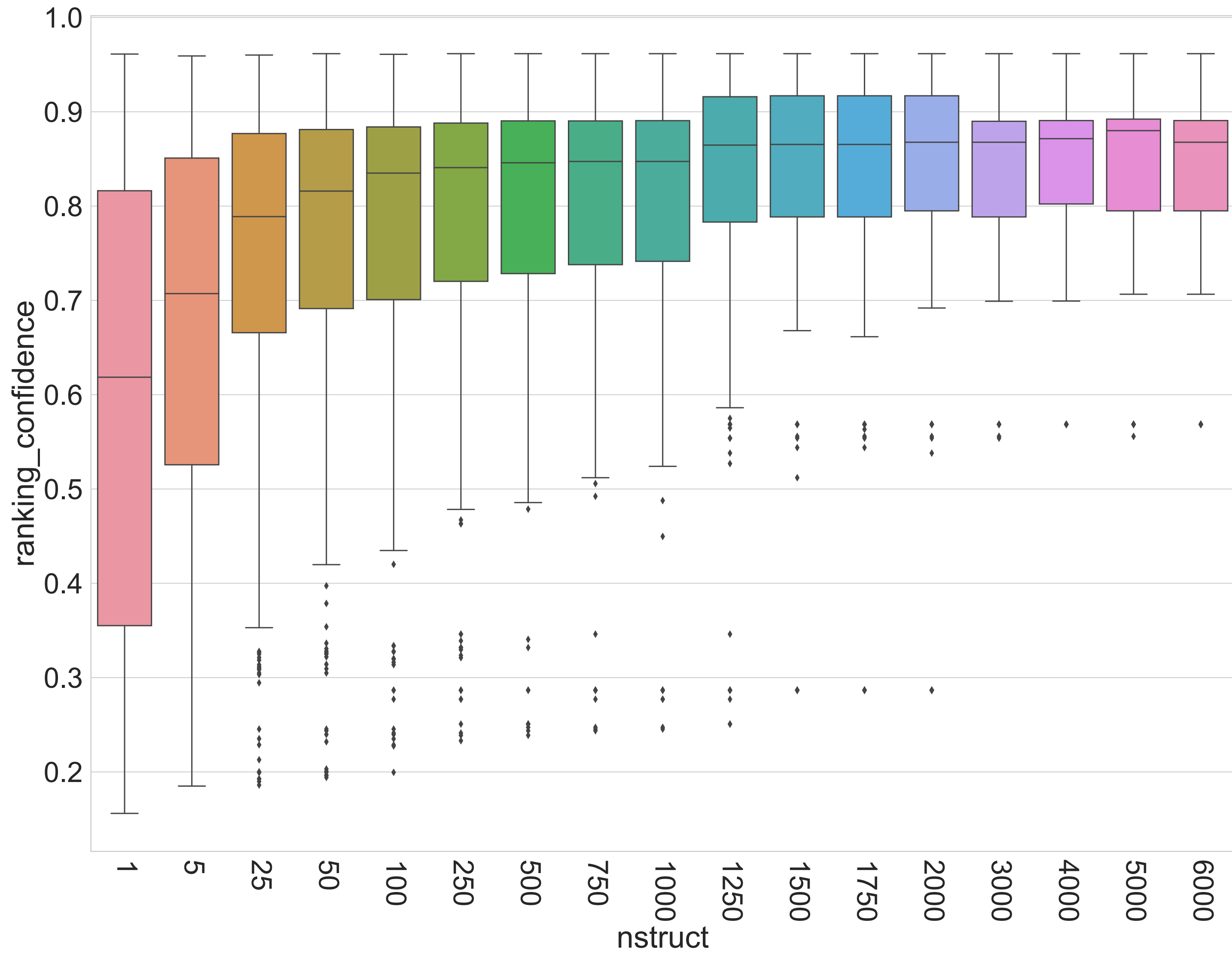## CASP15 results for Protein-Protein interactions
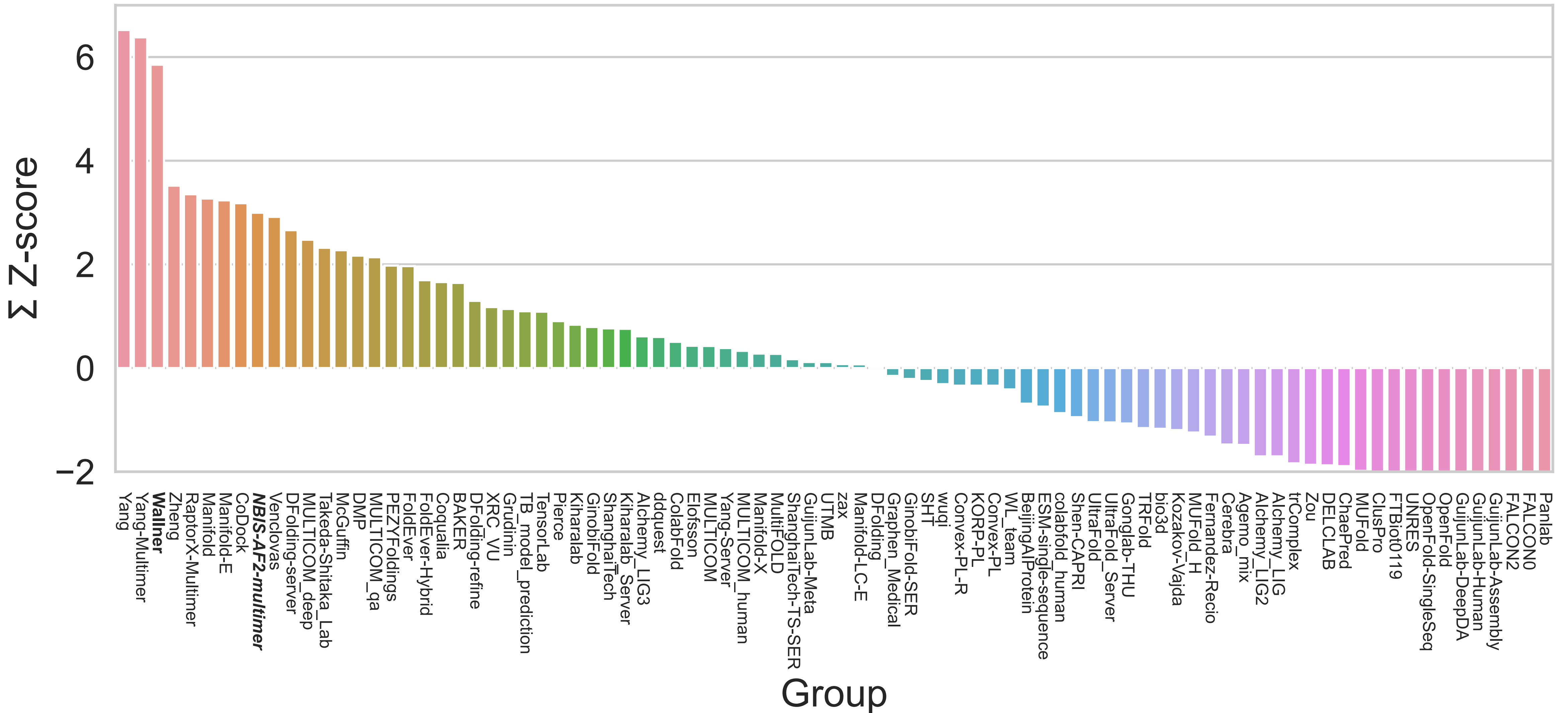
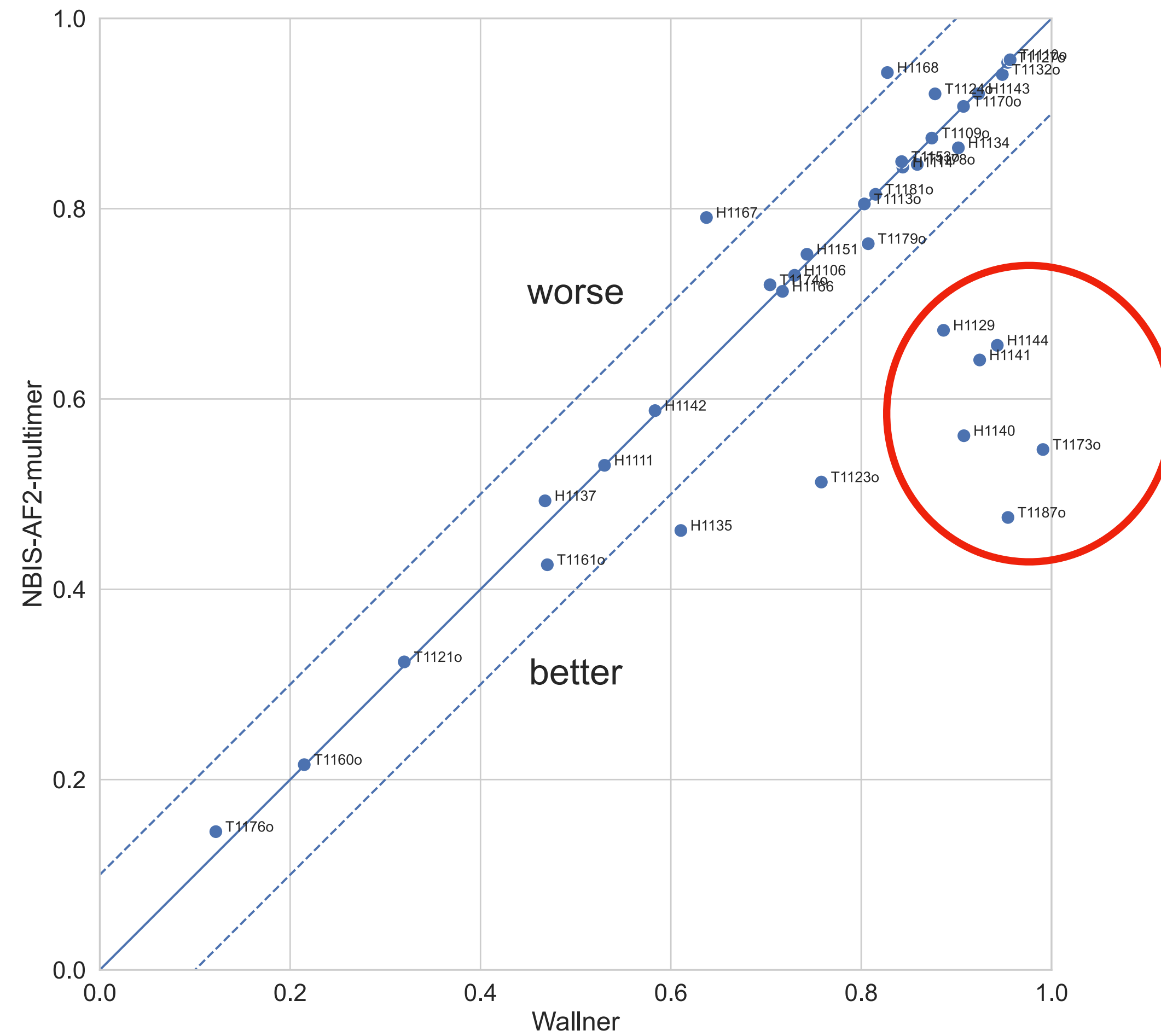# Comparison to baseline
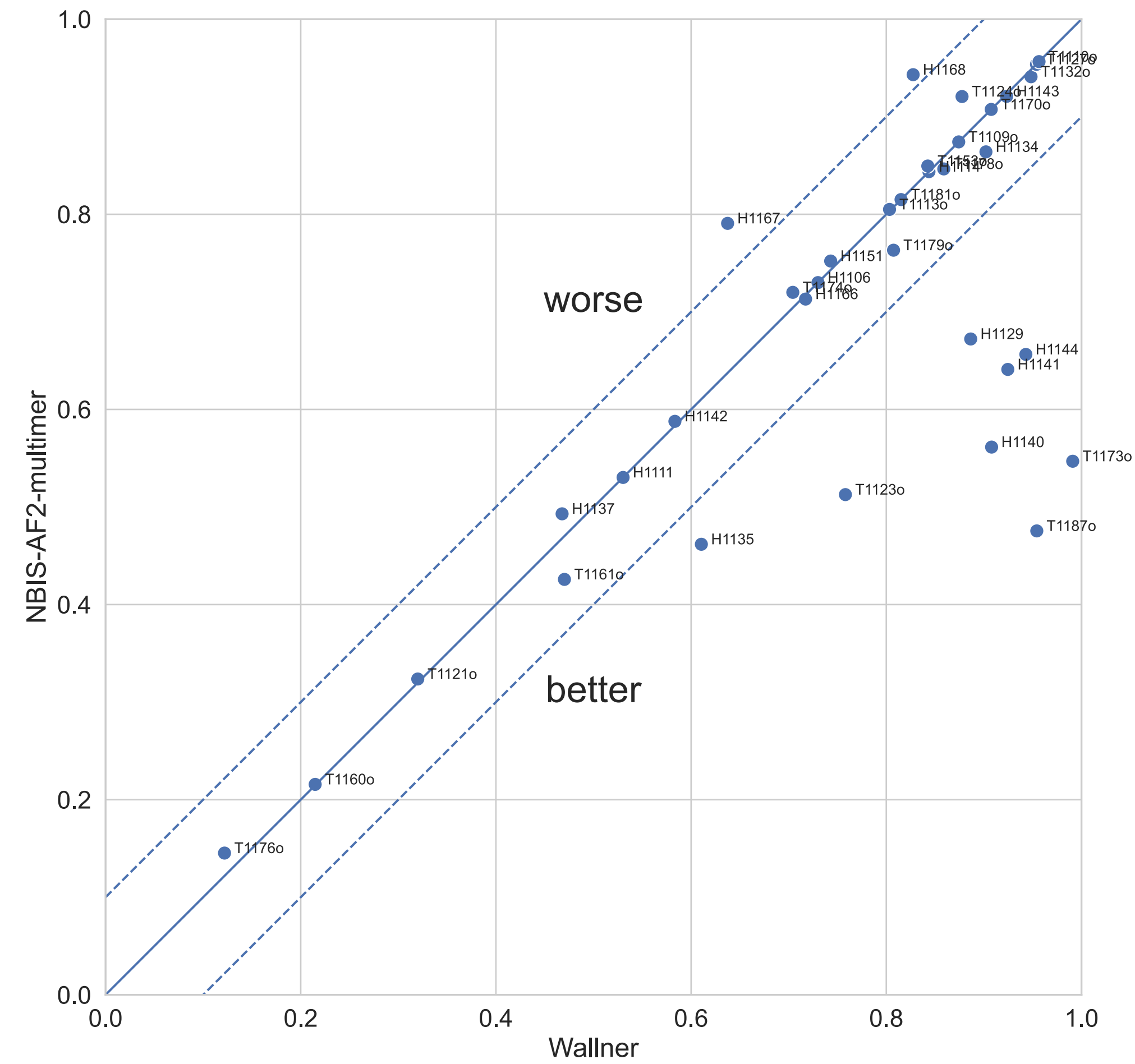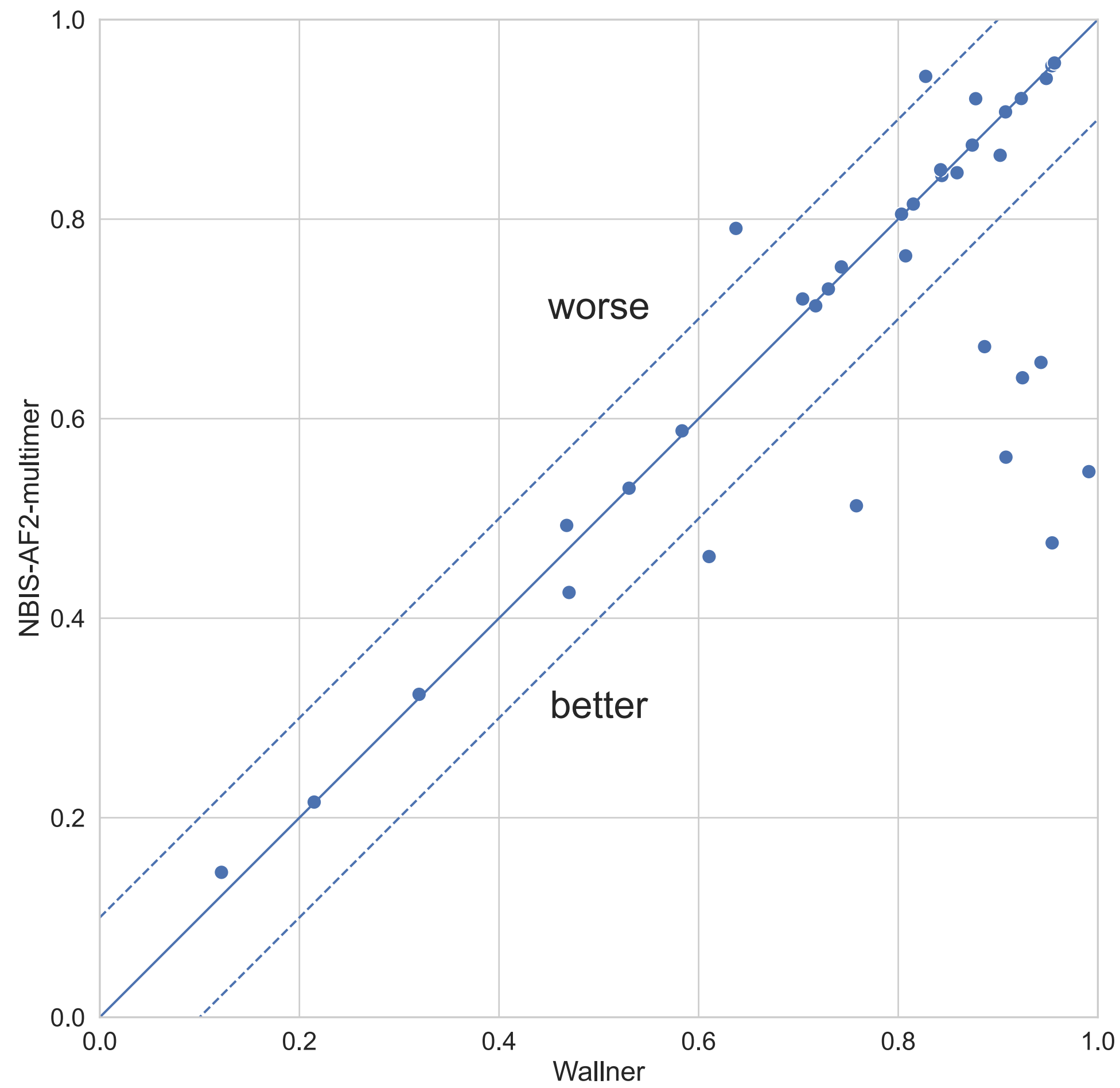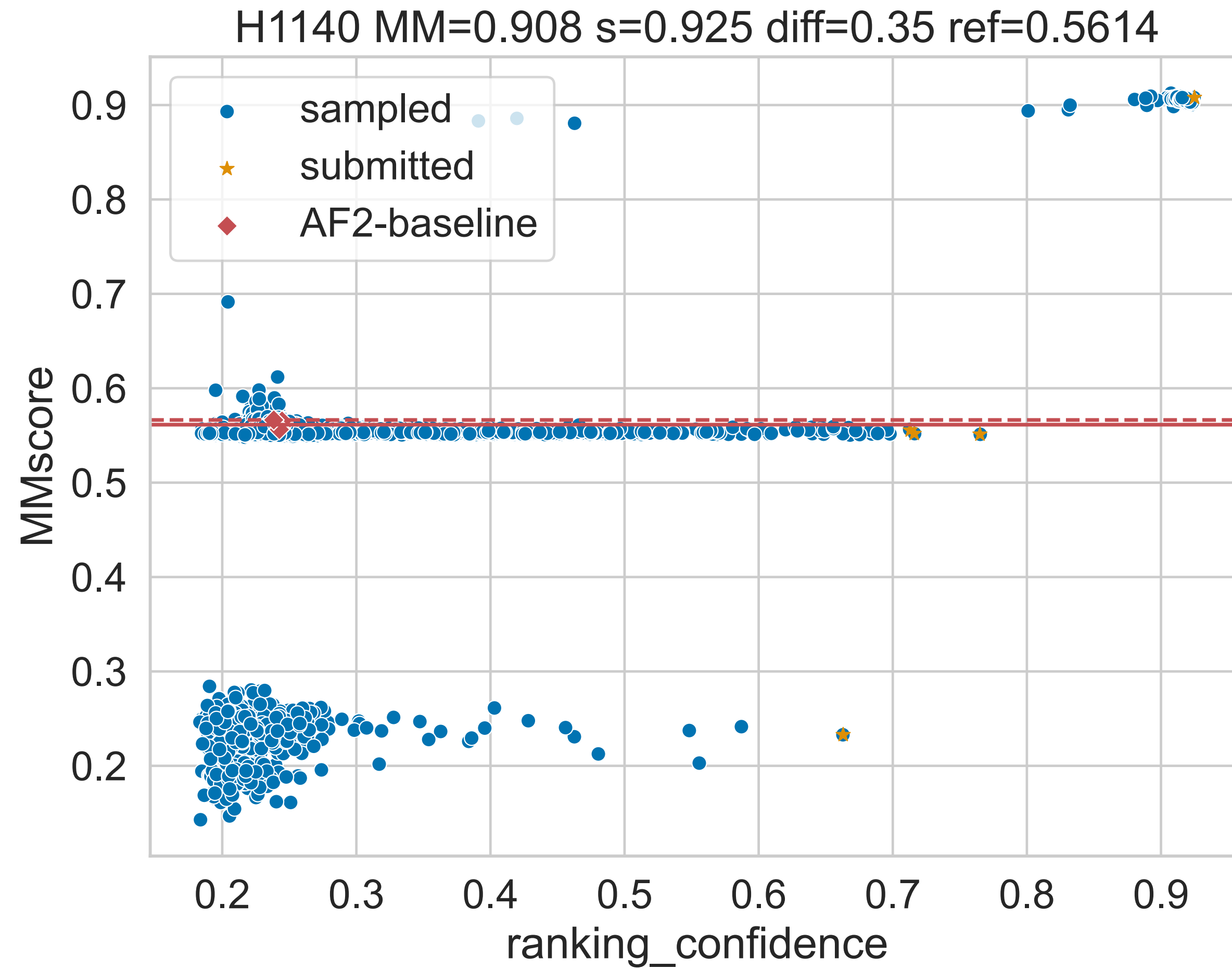
Sampling improves model quality
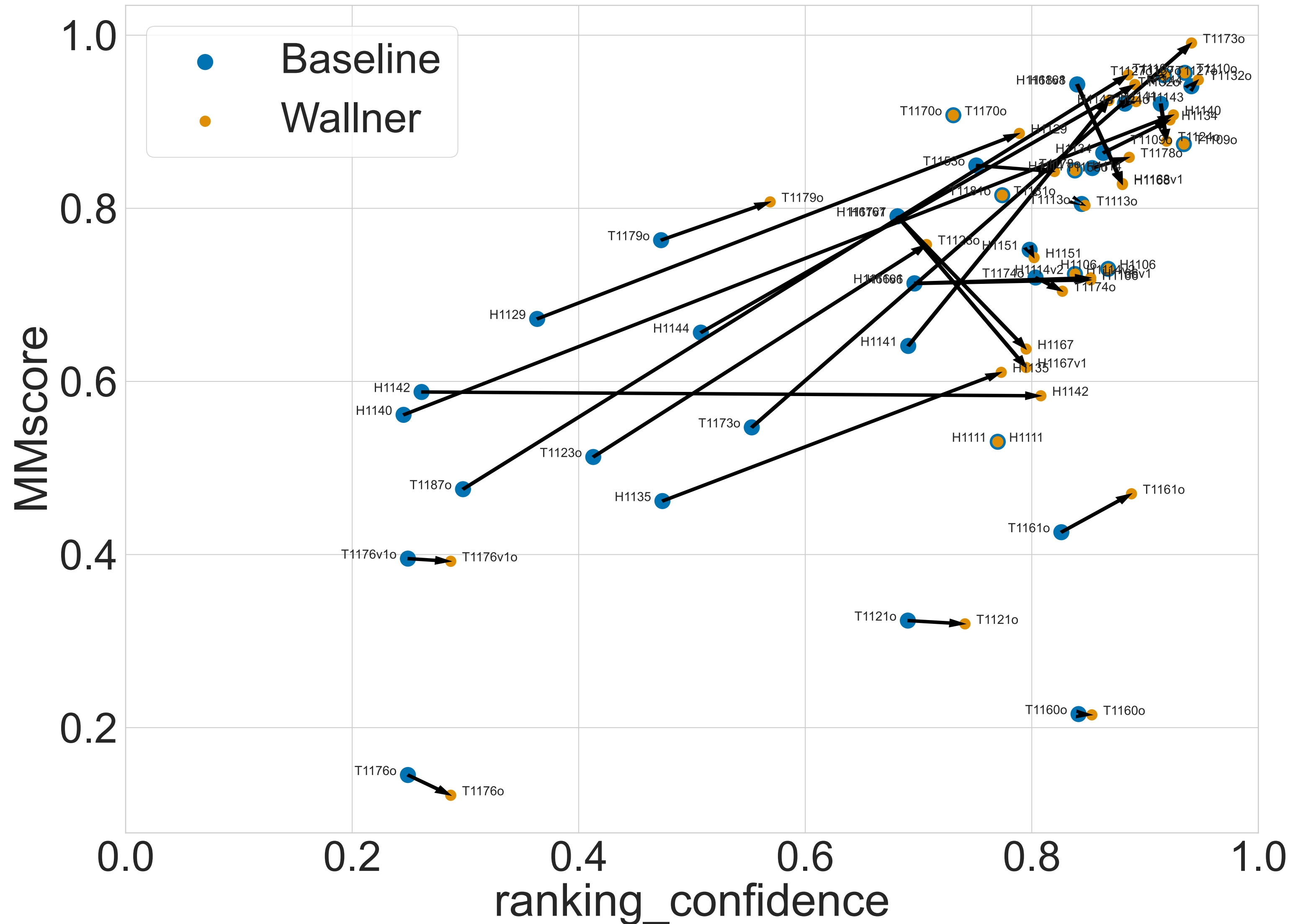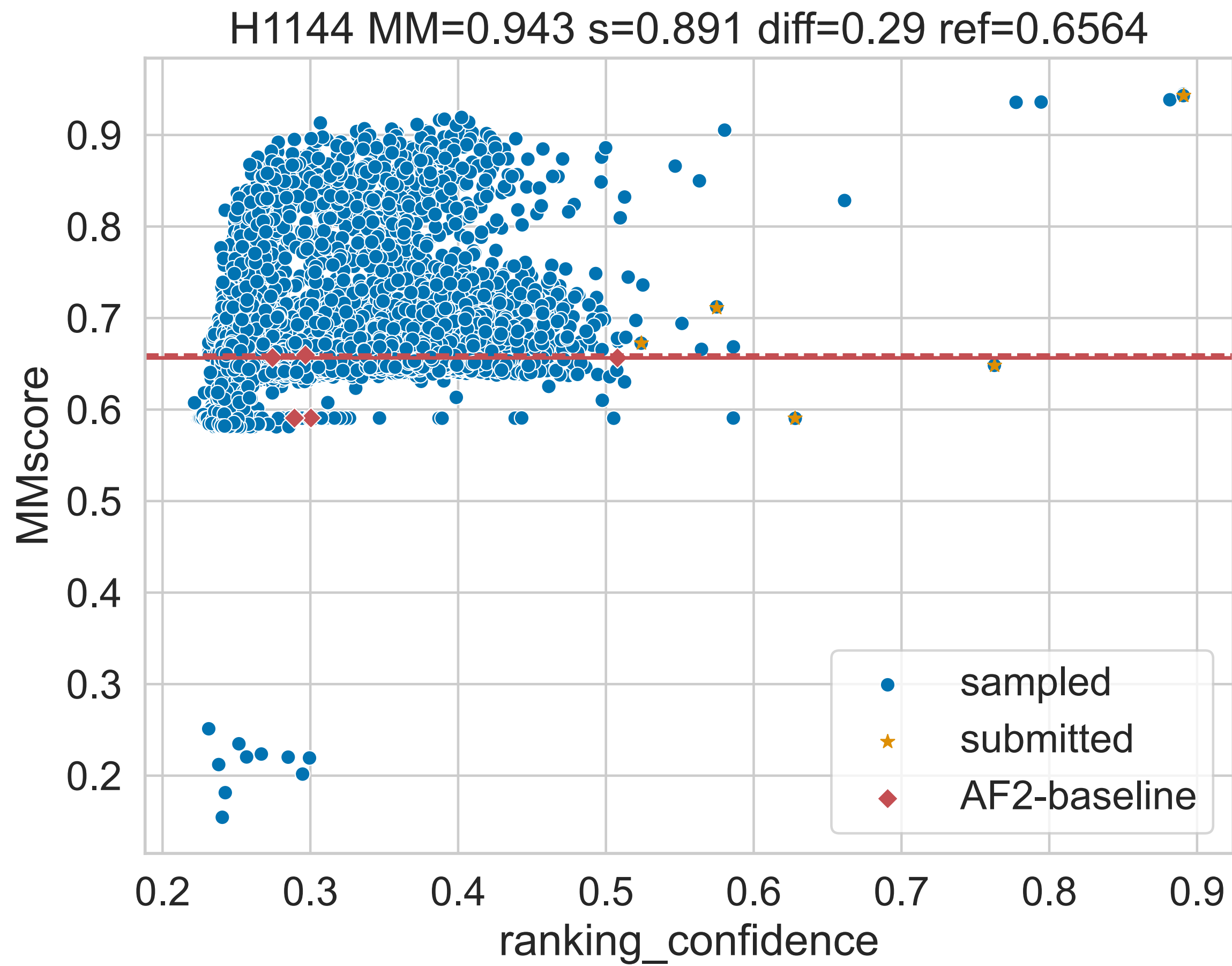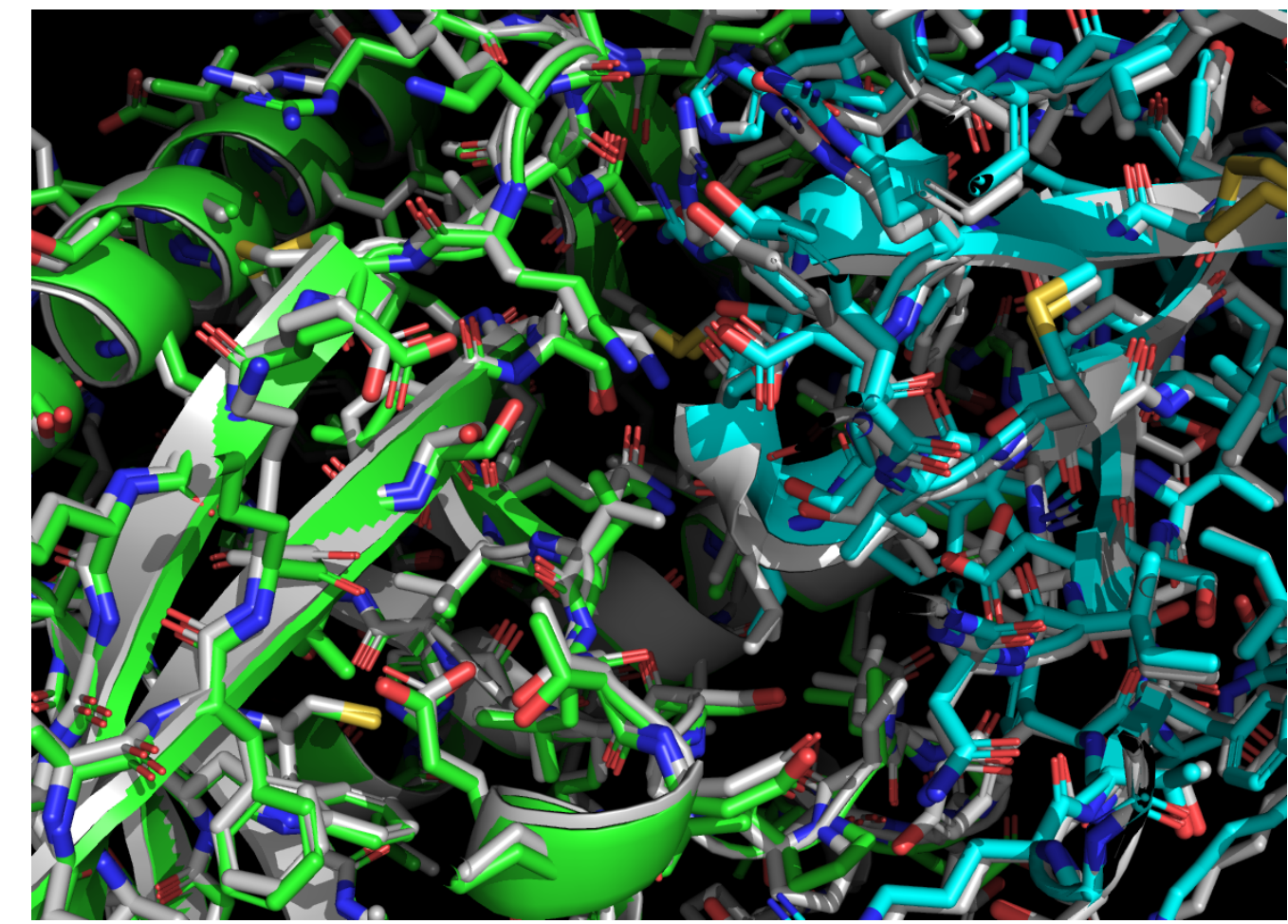
How much sampling is needed?
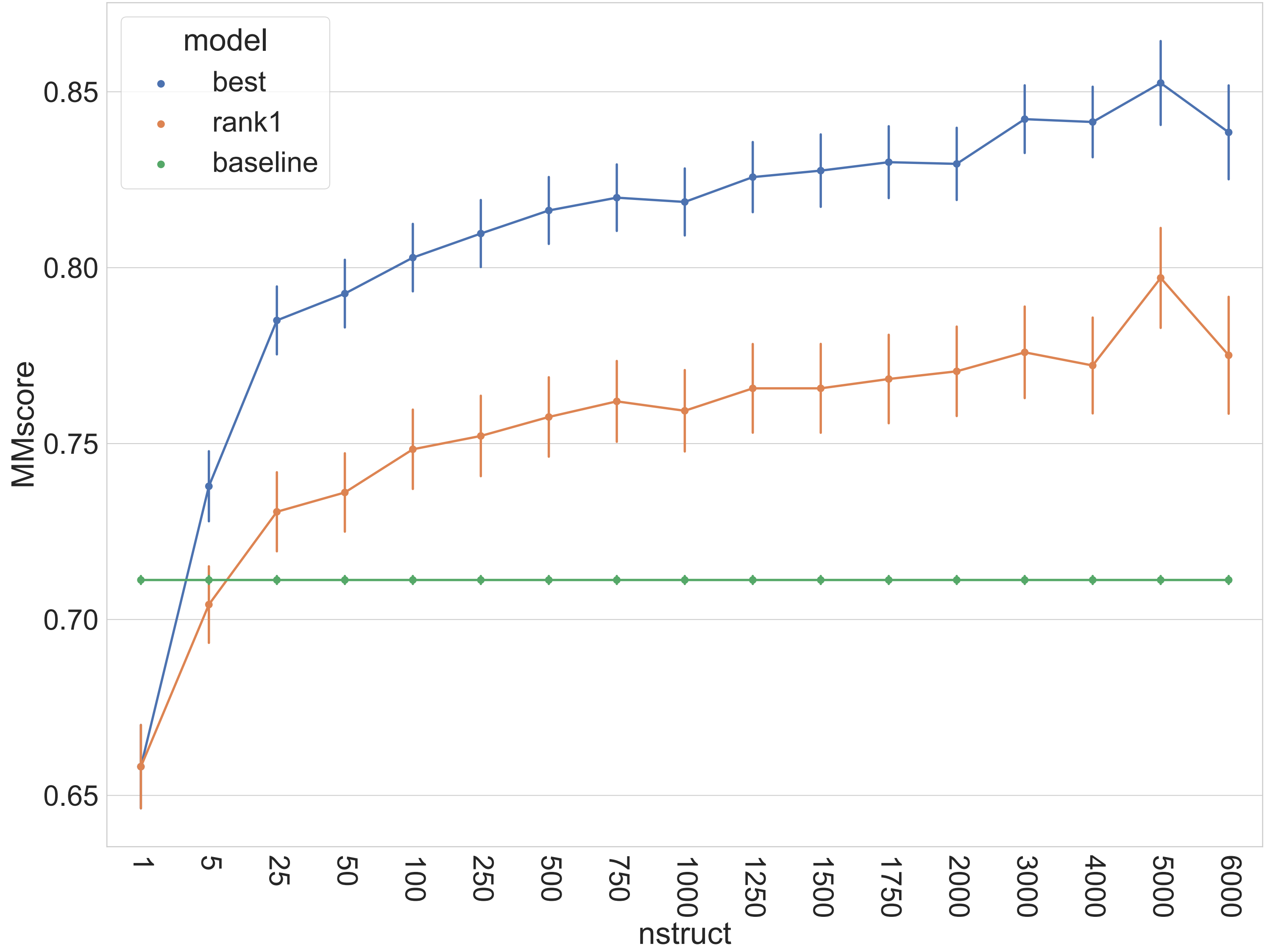
Ranking score improves

What went right

H1140

H1140 MM=0.908 s=0.925 diff=0.35 ref=0.5614

H1144

H1144 MM=0.943 s=0.891 diff=0.29 ref=0.6564

H1129

T1187o

T1187o MM=0.954 s=0.885 diff=0.48 ref=0.4755
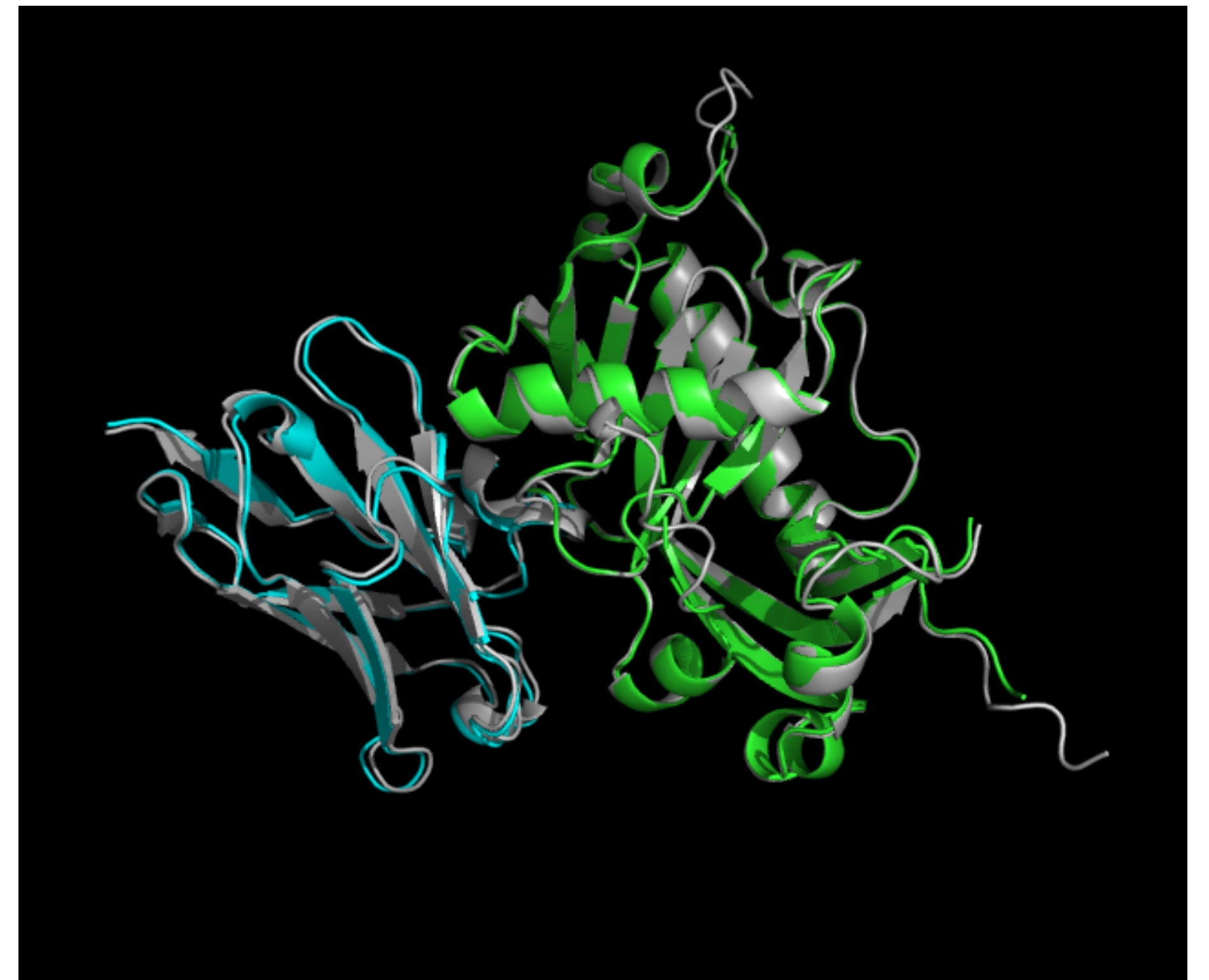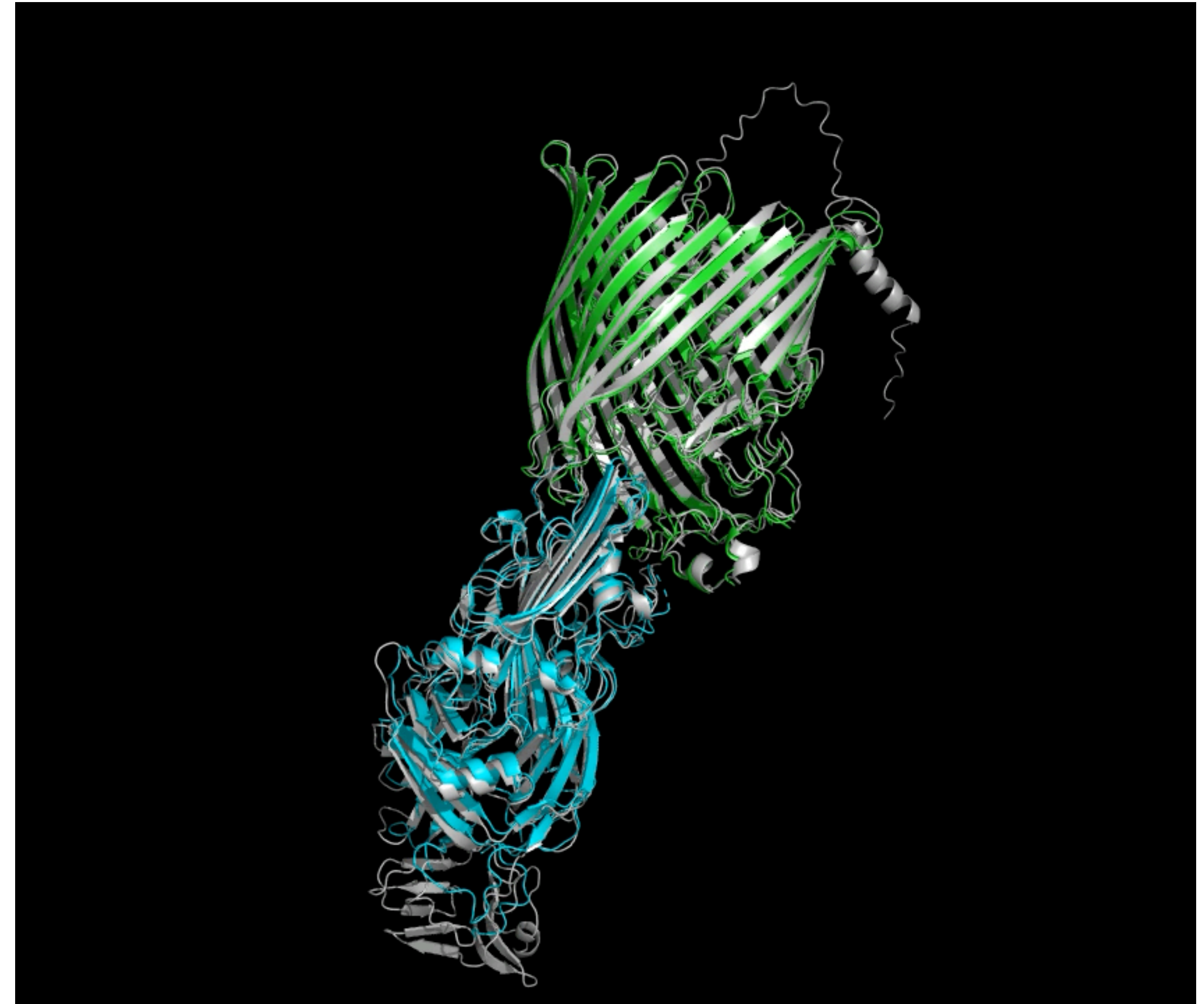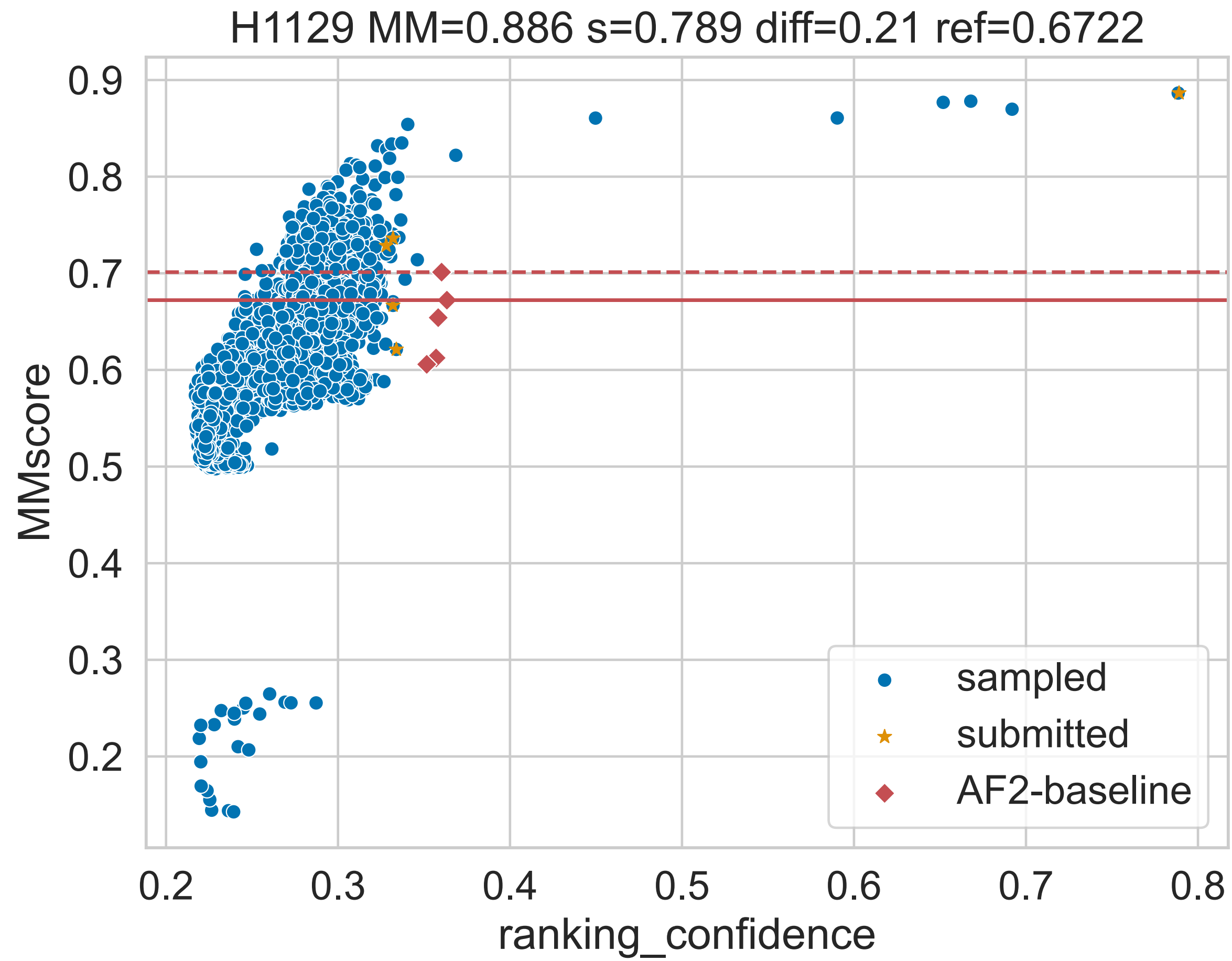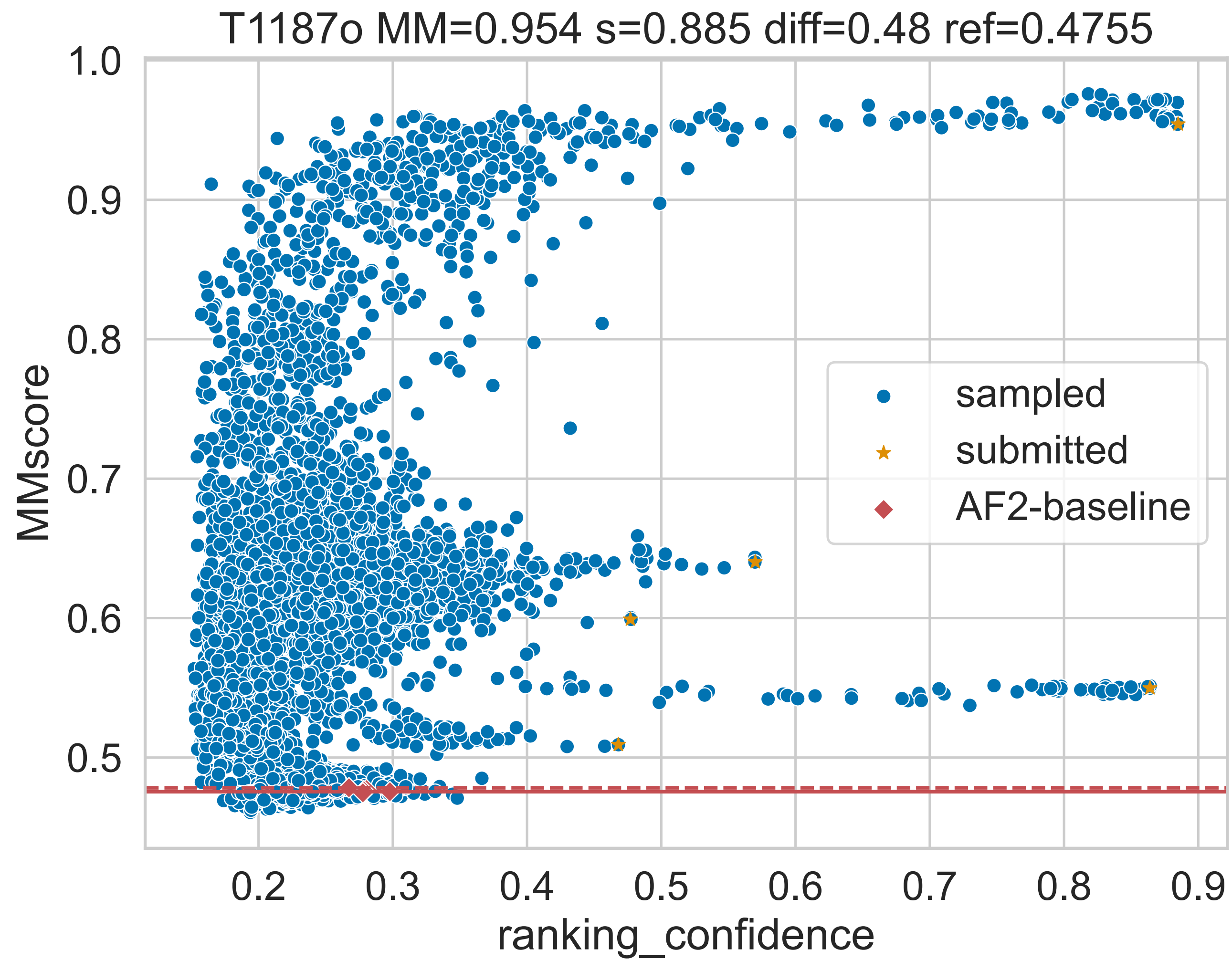
T1161, intertwined dimer

T1161o MM=0.470 s=0.888 diff=0.04 ref=0.4258

Native
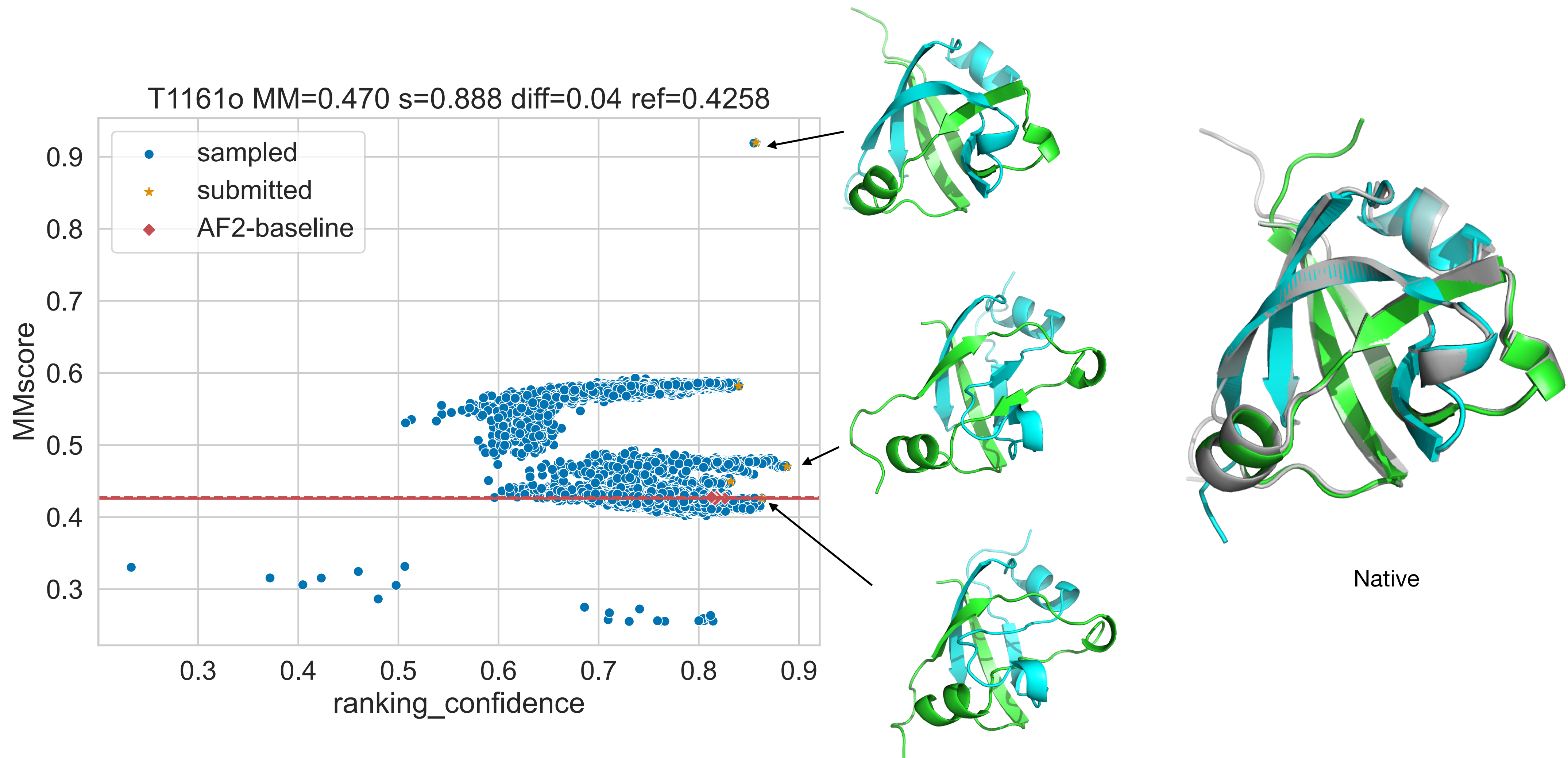
What went wrong

Baseline selects better models

H1168 MM=0.828 s=0.880 diff=-0.12 ref=0.9431

Baseline selects better models

H1167 MM=0.637 s=0.795 diff=-0.15 ref=0.7907

Legend:
- sampled
- submitted
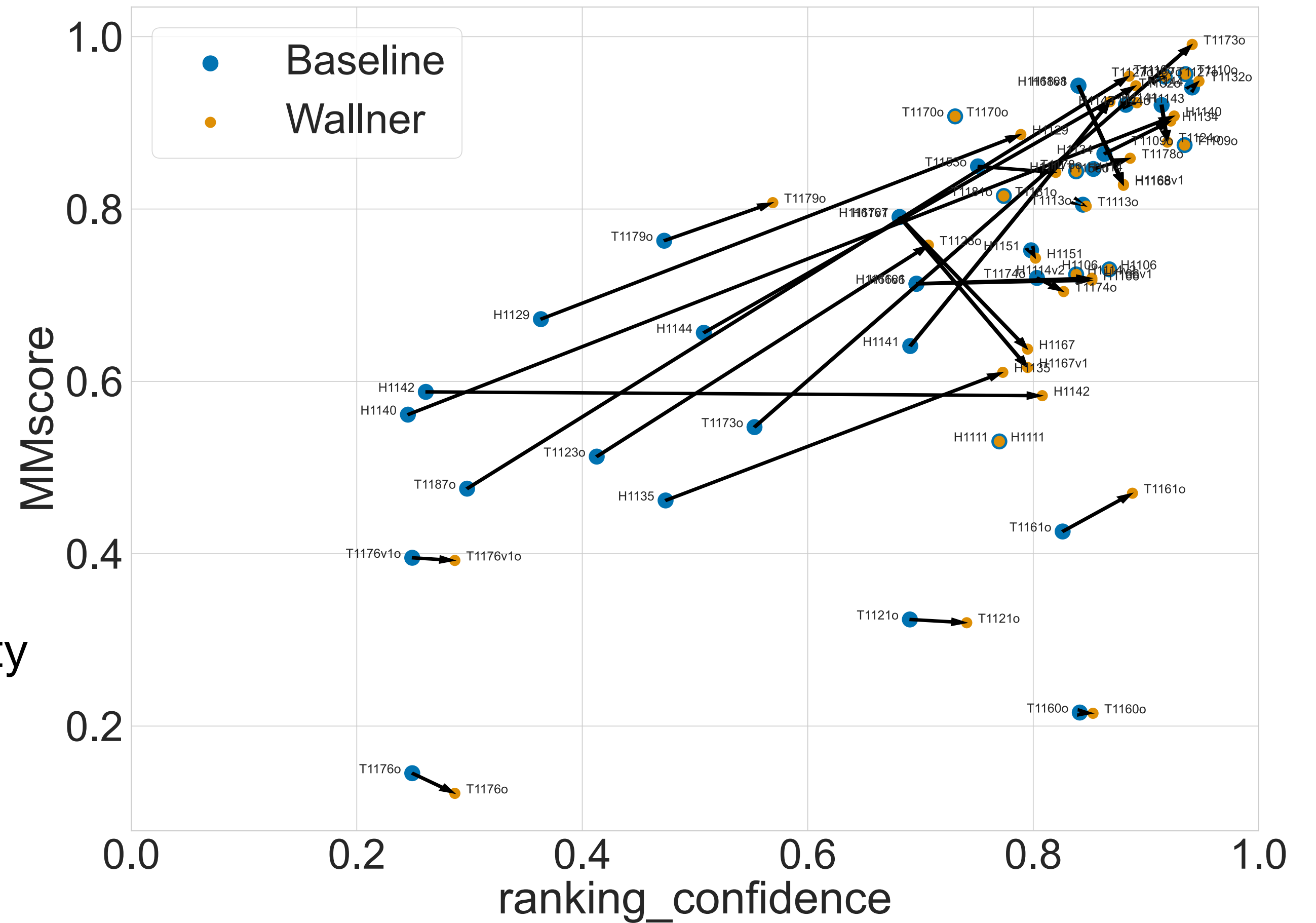- AF2-baseline

Native
Baseline
Wallner
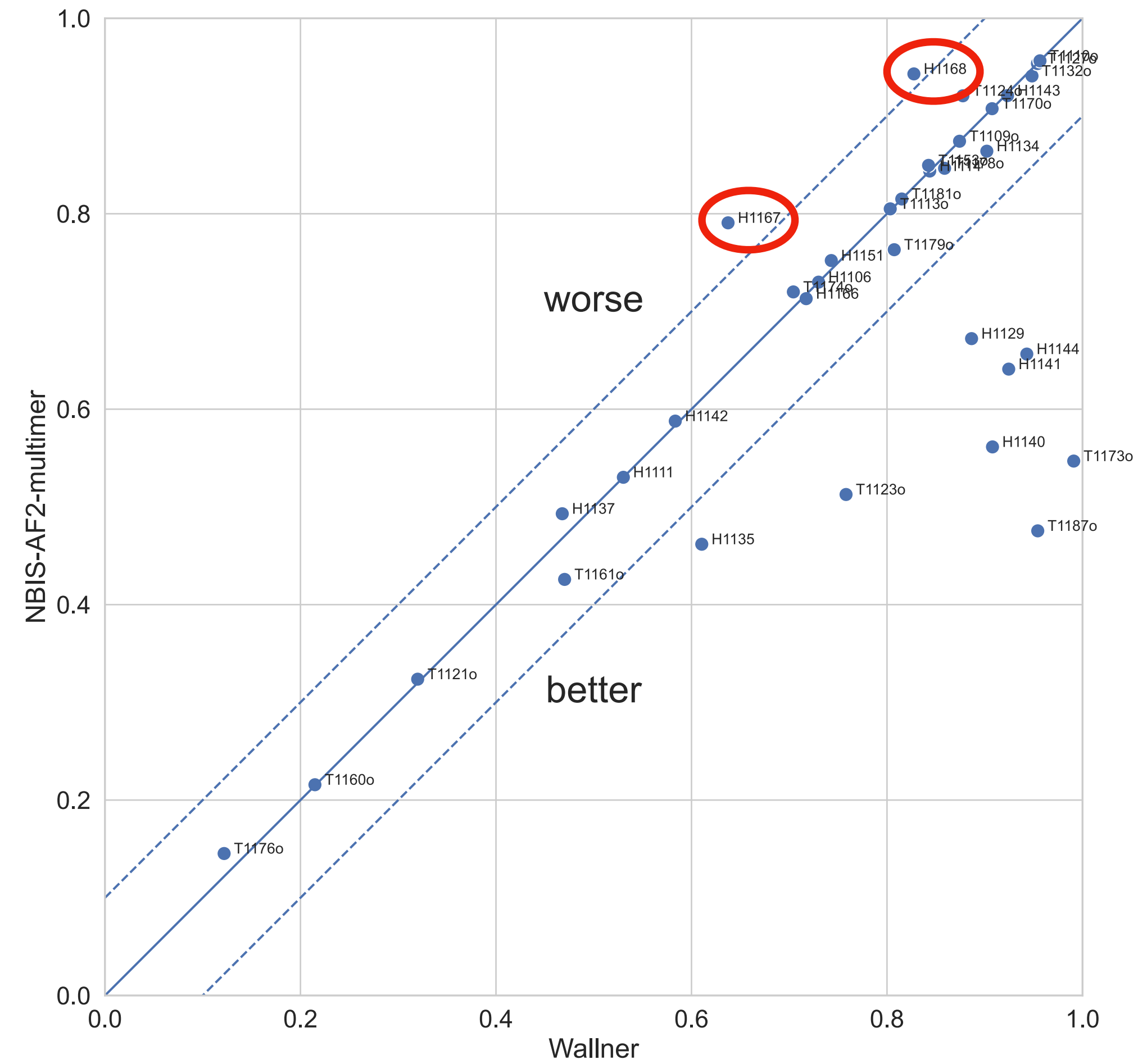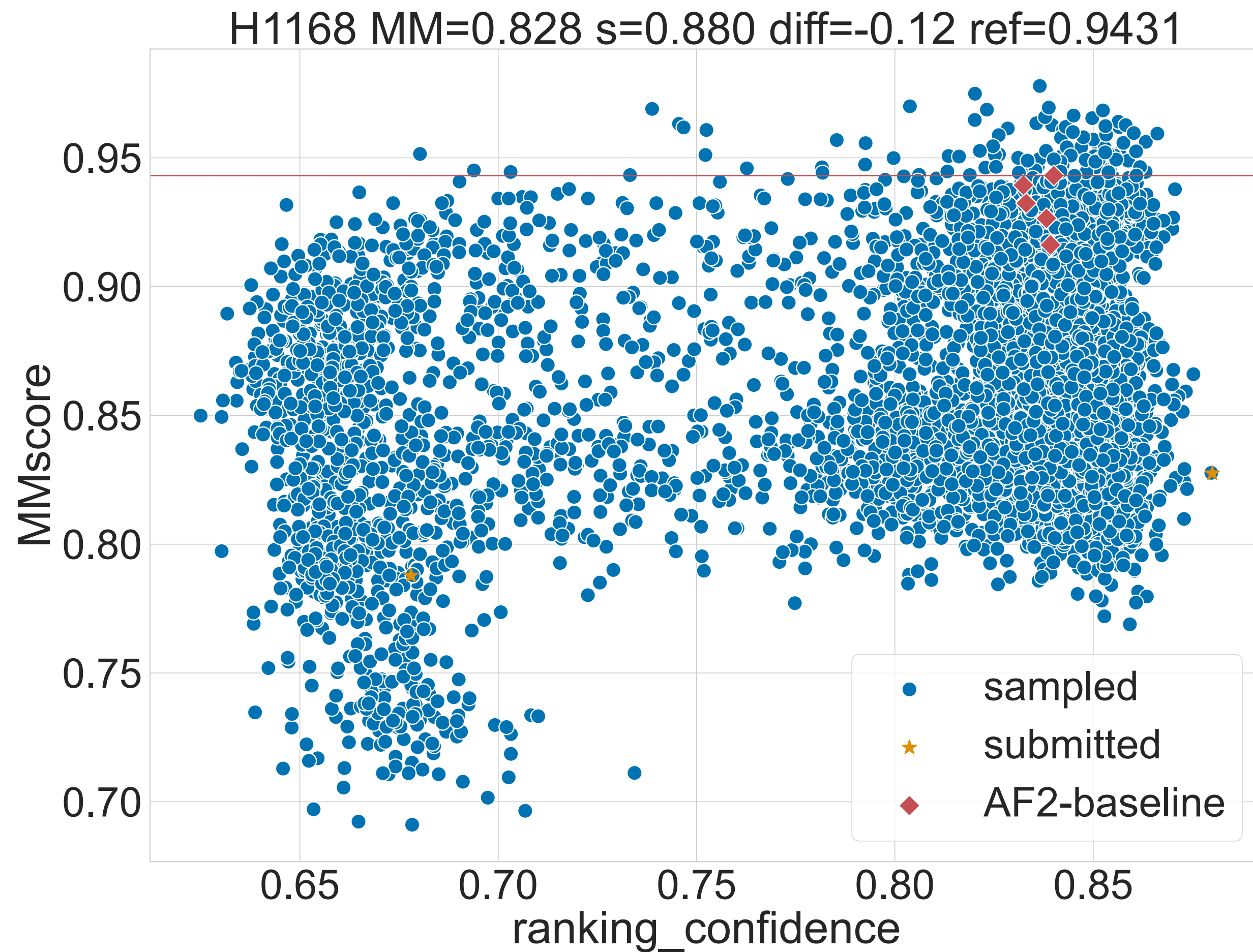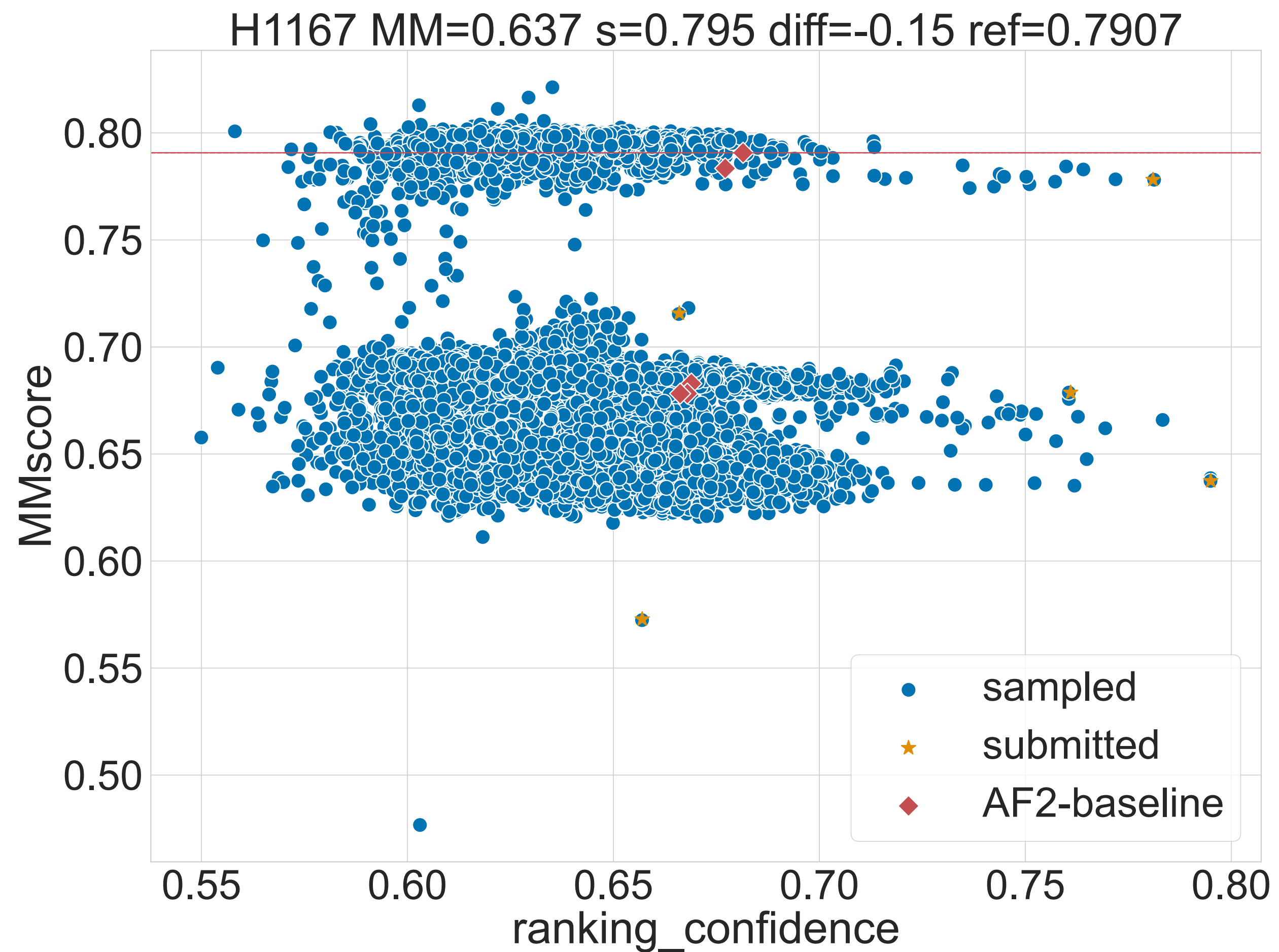
Summary

Sampling improves AlphaFold predictions.

- Increase

    - the number of models

    - recycles

- get **more diversity** using

    - Dropout

- Potential improvements: get even more diversity

    - Increase dropout rate

    - subsample MSA

    - Dropout+subsample MSA

Acknowledgement

Isak Johansson-Åkhe

frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fbinf.2022.959160

Check for updates

OPEN ACCESS

EDITED BY
Dinler Amaral Antunes,
University of Houston, United States

Improving peptide-protein docking with AlphaFold-Multimer using forced sampling

Isak Johansson-Åkhe and Björn Wallner*

Division of Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden