Evaluating and refining EM structures with TEMPy

1. Methodology
2. Correlation between scores
3. Examples (for each: show it improves, show SMOC, show worst)
    a. T1036s1: Glycoprotein B
    b. T1092/T1096: Polymerase
    c. T1099: Duck Hepatitis B Virion

Methodology

Compute scores against experimental map

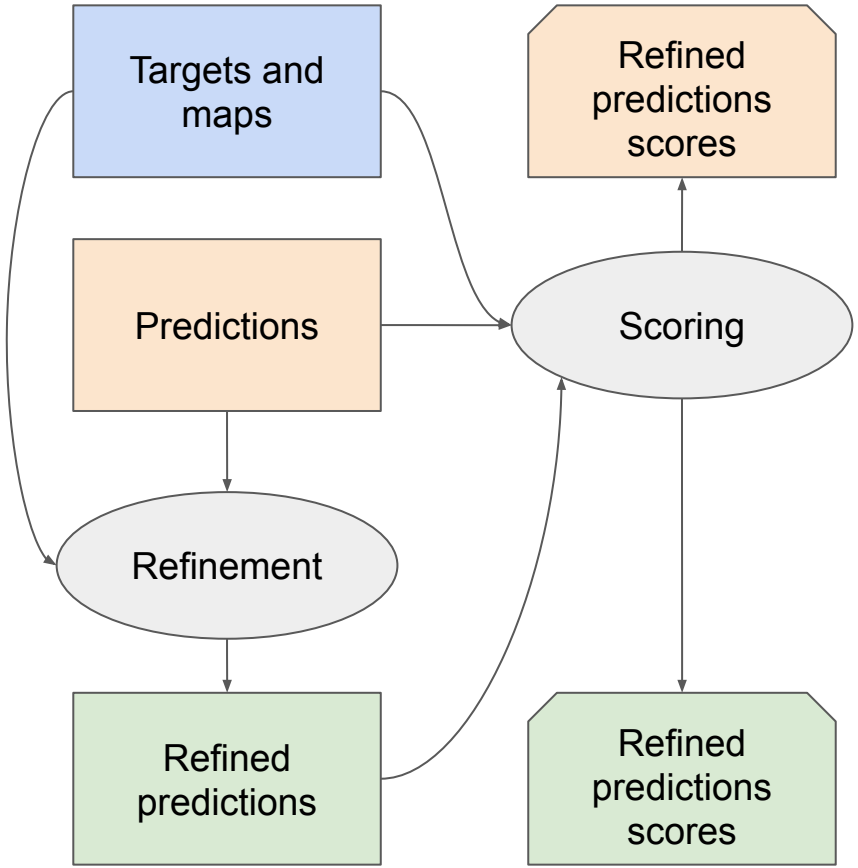
- **C**ross-**C**orrelation **C**oefficient
- **N**ormalised **M**utual **I**nformation
- **S**egment-based **M**anders **O**verlap **C**oefficient

Refinement of top predictions

- Based on map
- Maintaining stereochemistry
- Automated

Then recompute the scores

Python code with TEMPy, Openmm, numpy, pandas, matplotlib
3D rendering with Chimera/ChimeraX

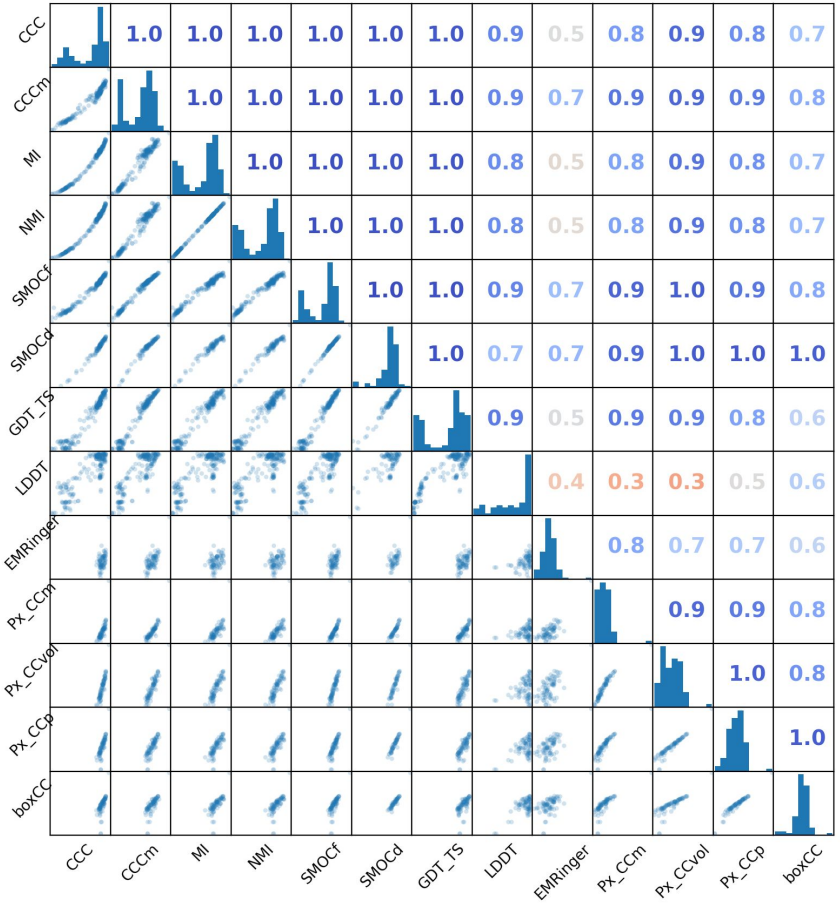


Scoring correlation

All scores exhibit a relatively high degree of correlation.

Strong correlations with GDT_TS, weaker correlations for LDDT.

Matrix for T1036s1

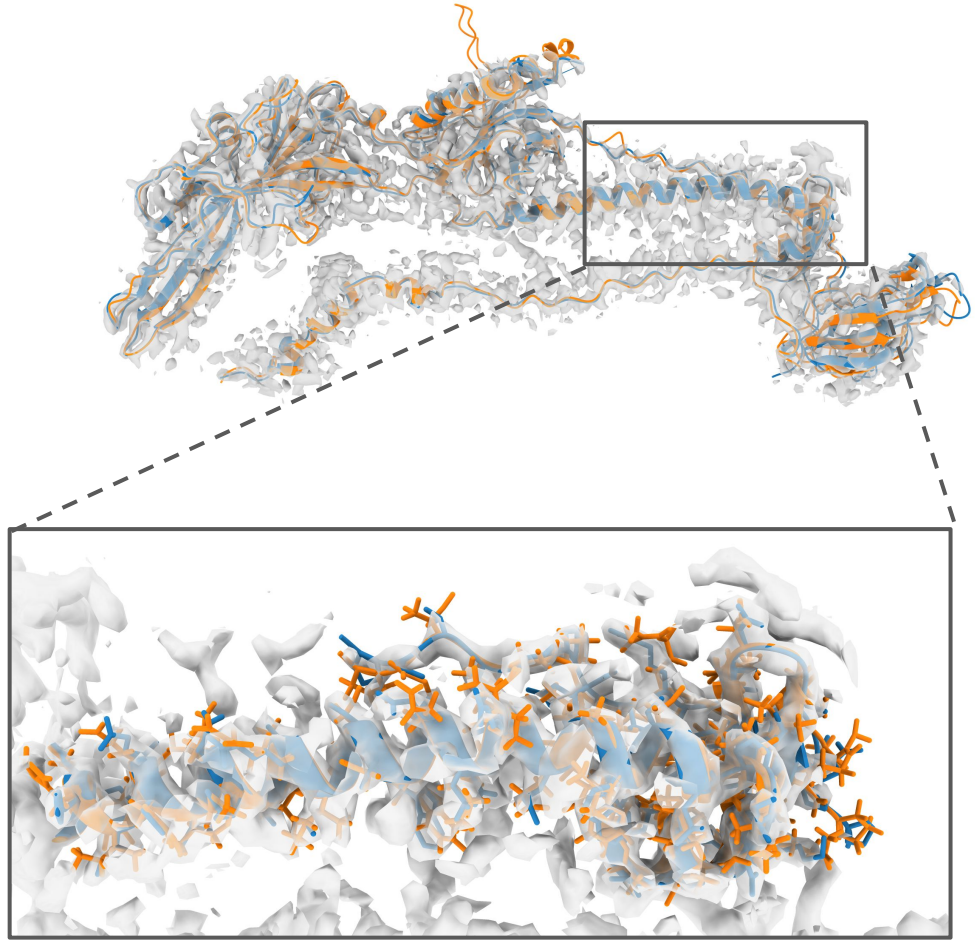


T1036s1: Glycoprotein B

TBM-easy
Resolution ~ 2.8 Å

RaptorX (487) is ranked #1, by GDT_TS and CCC.

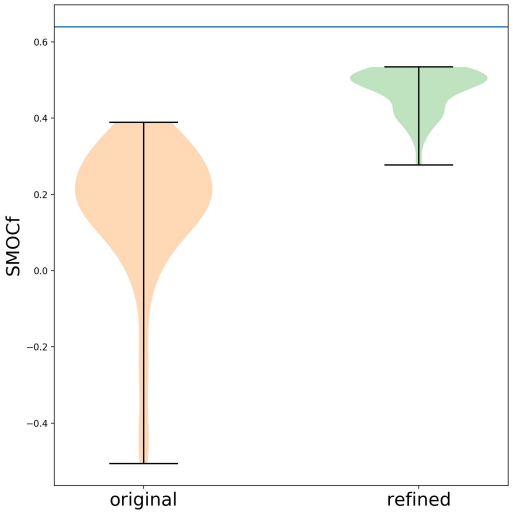
The fold is correct,
and side-chains are mostly right.

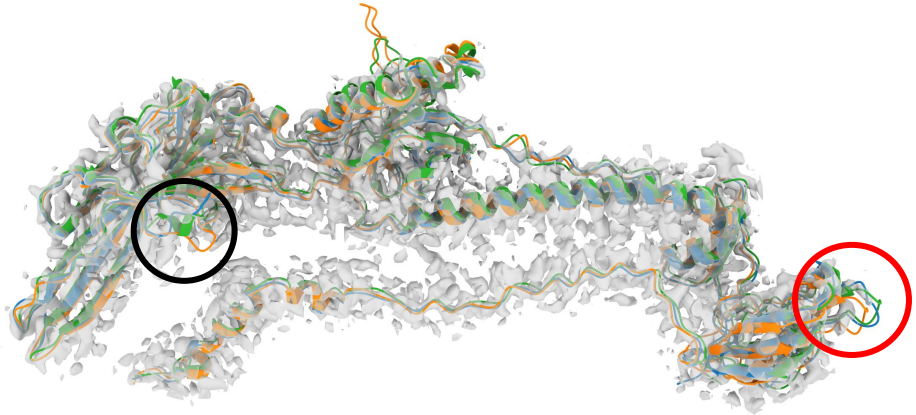


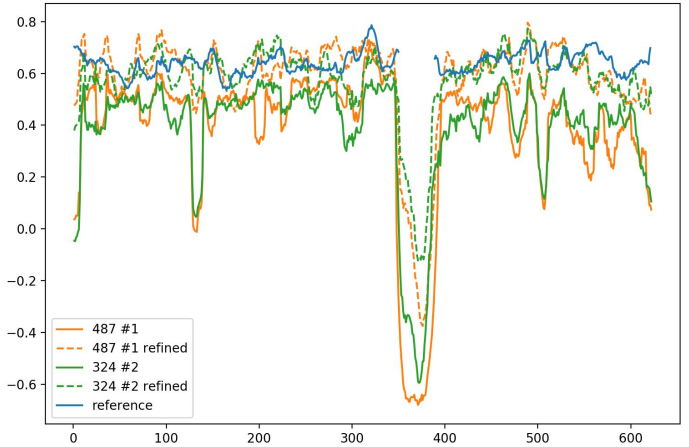
T1036s1: Glycoprotein B

TBM-easy
Resolution ~ 2.8 Å

Automatically refined models are almost as good as experimental reference!

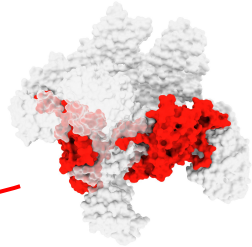




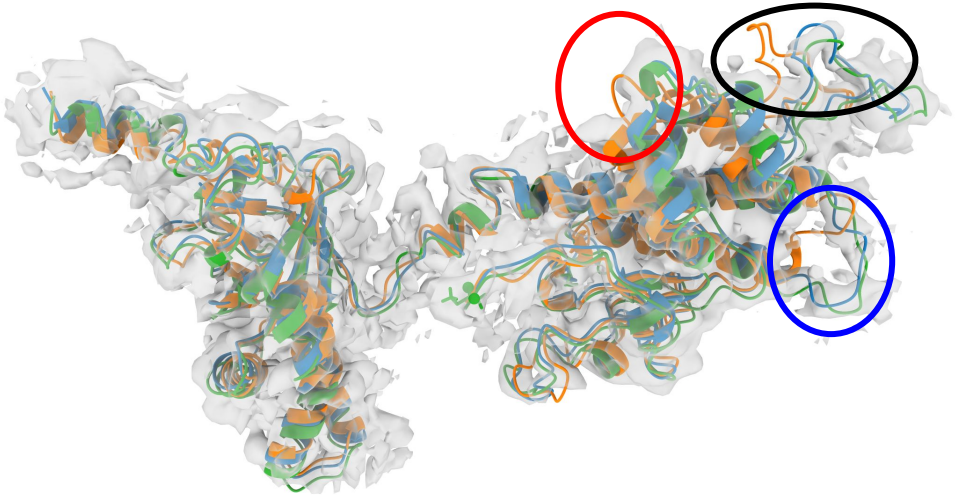


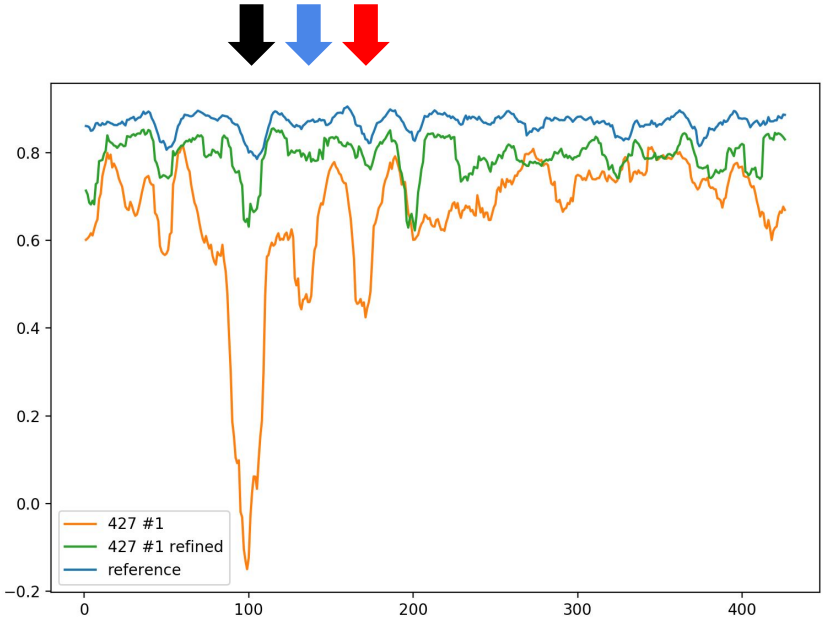
T1092: polymerase

TBM-easy
Resolution ~ 3.8 Å



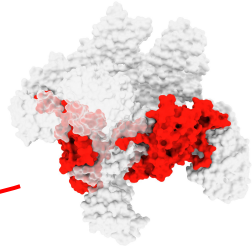
Experimental SMOC: ~0.9
Predicted SMOC: ~0.6
Refined SMOC: ~ 0.8

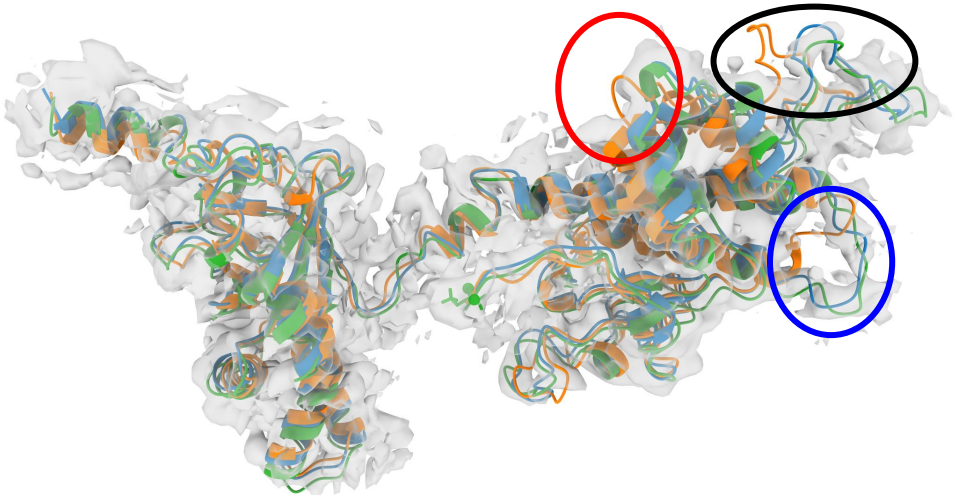




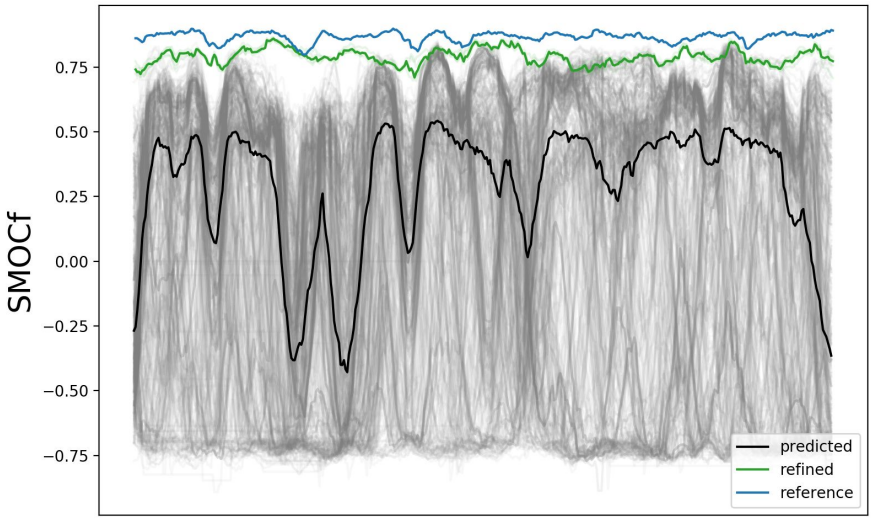
T1092: polymerase

TBM-easy
Resolution ~ 3.8 Å



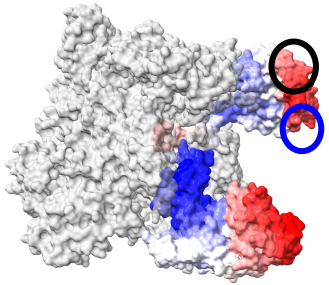


The SMOC score broadly correlates with the prediction quality.

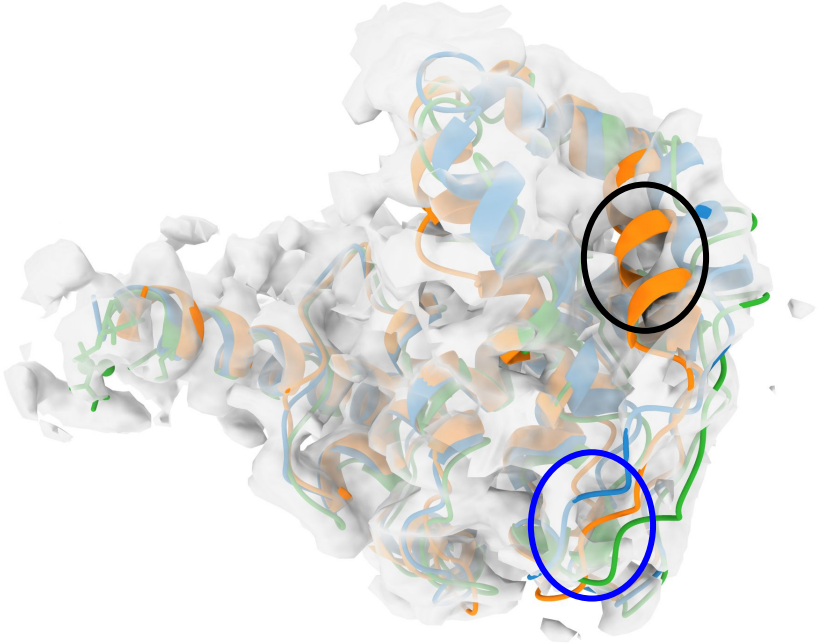


T1096-D2: polymerase

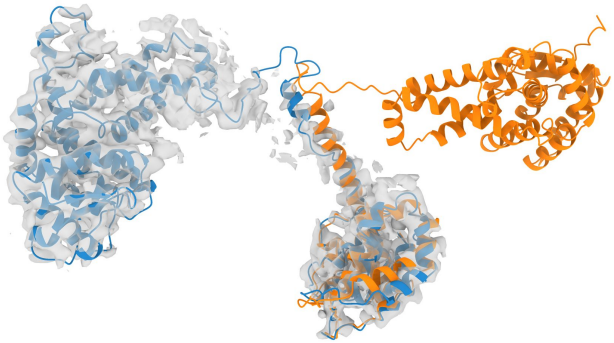
FM
Resolution ~ 3.8 Å



Low local resolution regions
(computed with Resmap)
were harder to predict (and have
lower SMOC scores)



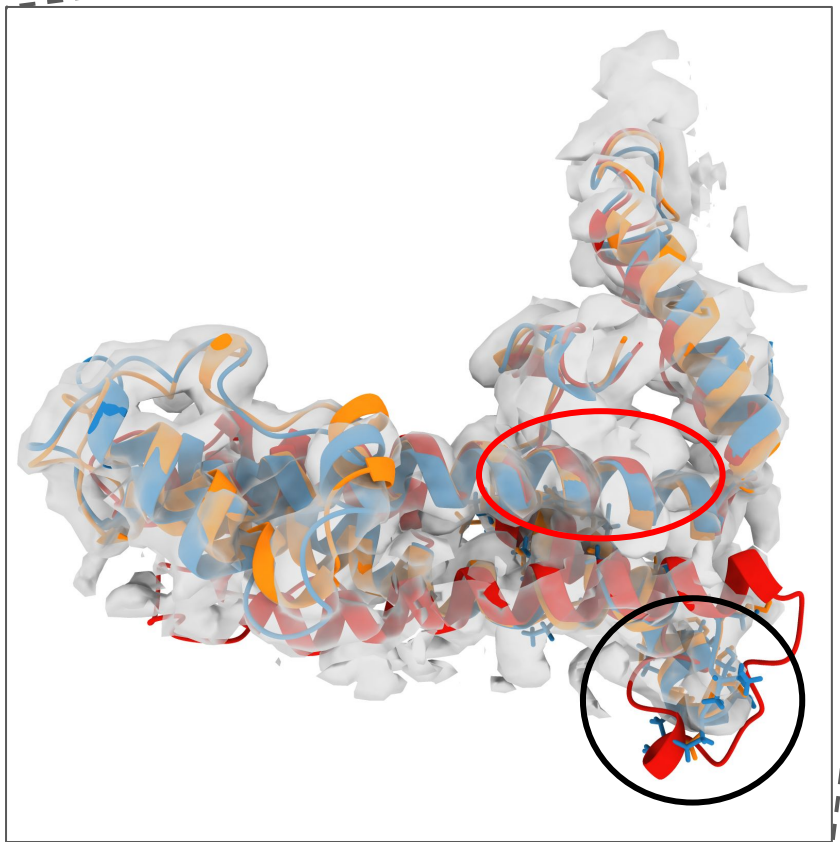
Both domains of T1096 were well
predicted, however the linker
(and domain orientation) was wrong!

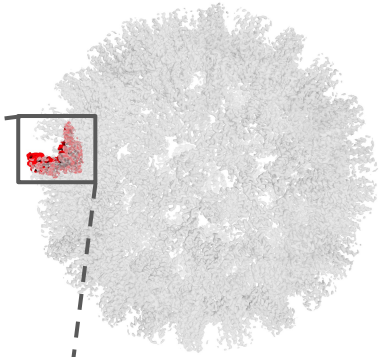


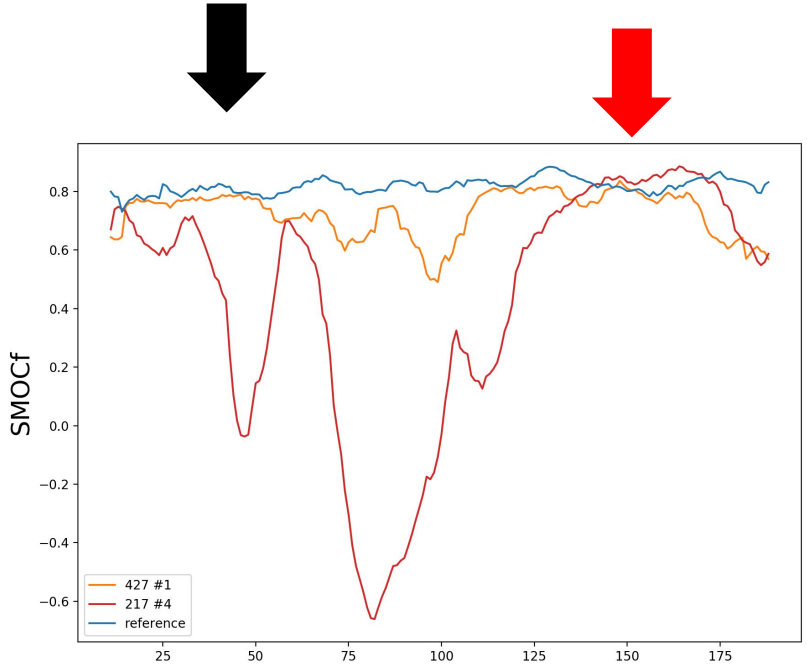
T1099: virus capsid

TBM-hard
Resolution ~ 3.7 Å

Some models have a better SMOC than the experimental reference, **before refinement**







Acknowledgments

Maya Topf

Andriy Kryshtafovych

Experimentalists and participants

Petr Leiman and Alec Fraser (UTMB)

Topf and Thalassinos groups

Sony Malhotra

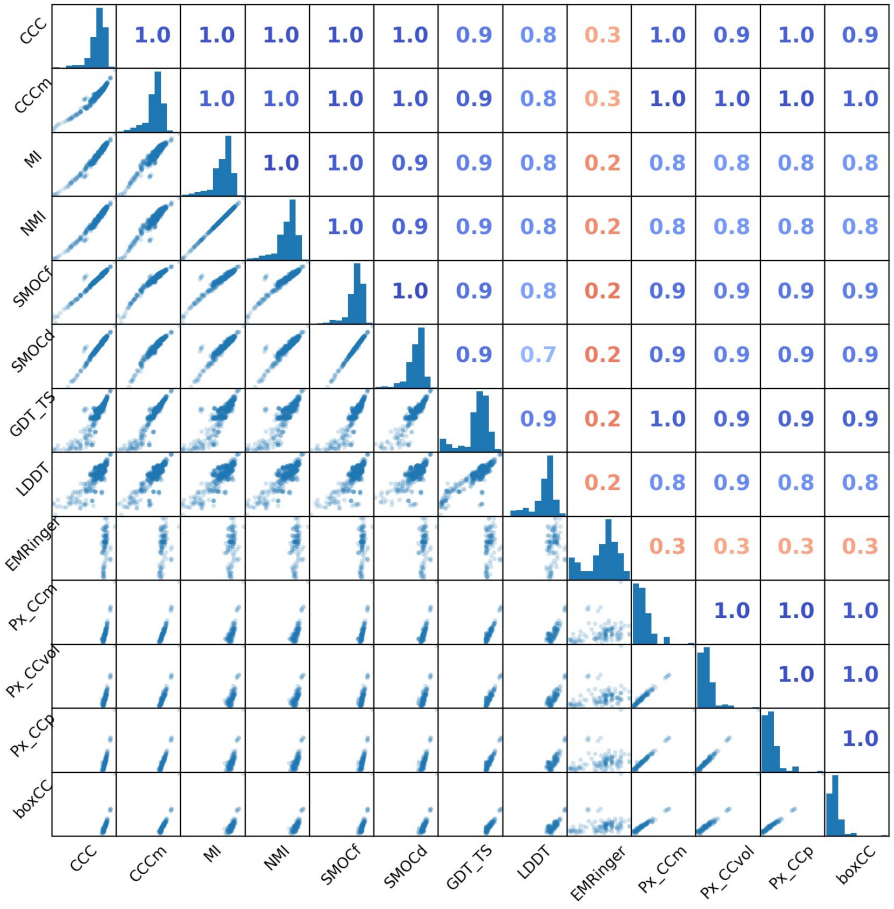
Tom Mulvaney

Agnel Praveen Joseph (STFC)

Annex

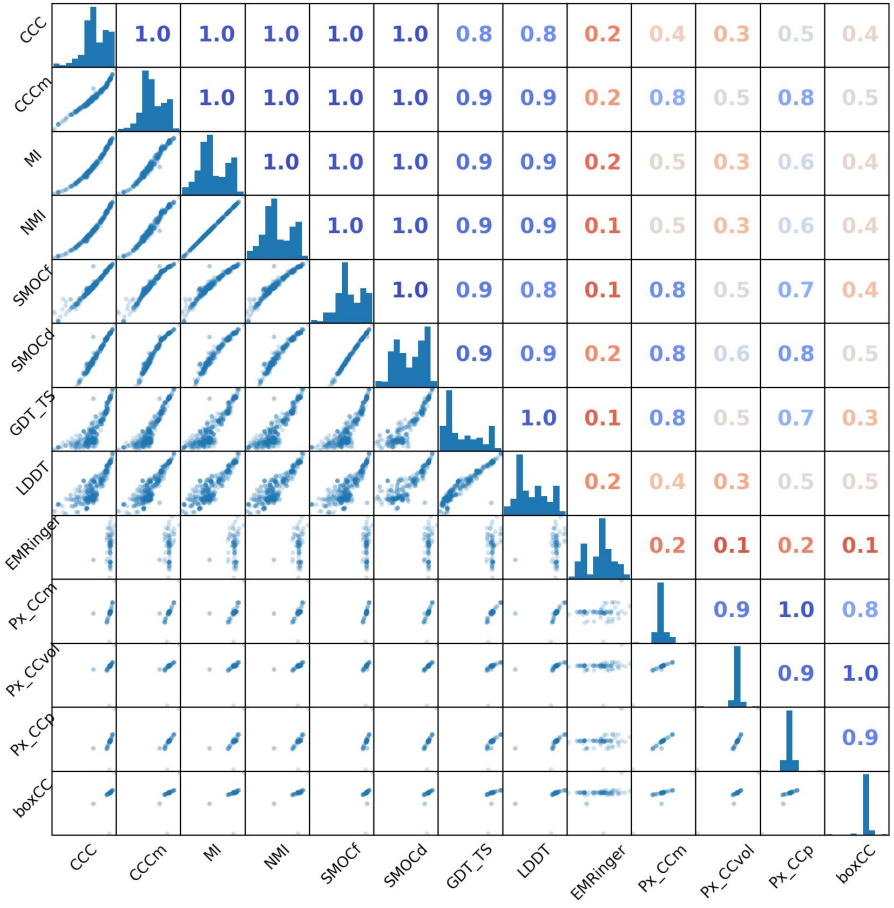
Scoring correlation

For T1092-D2



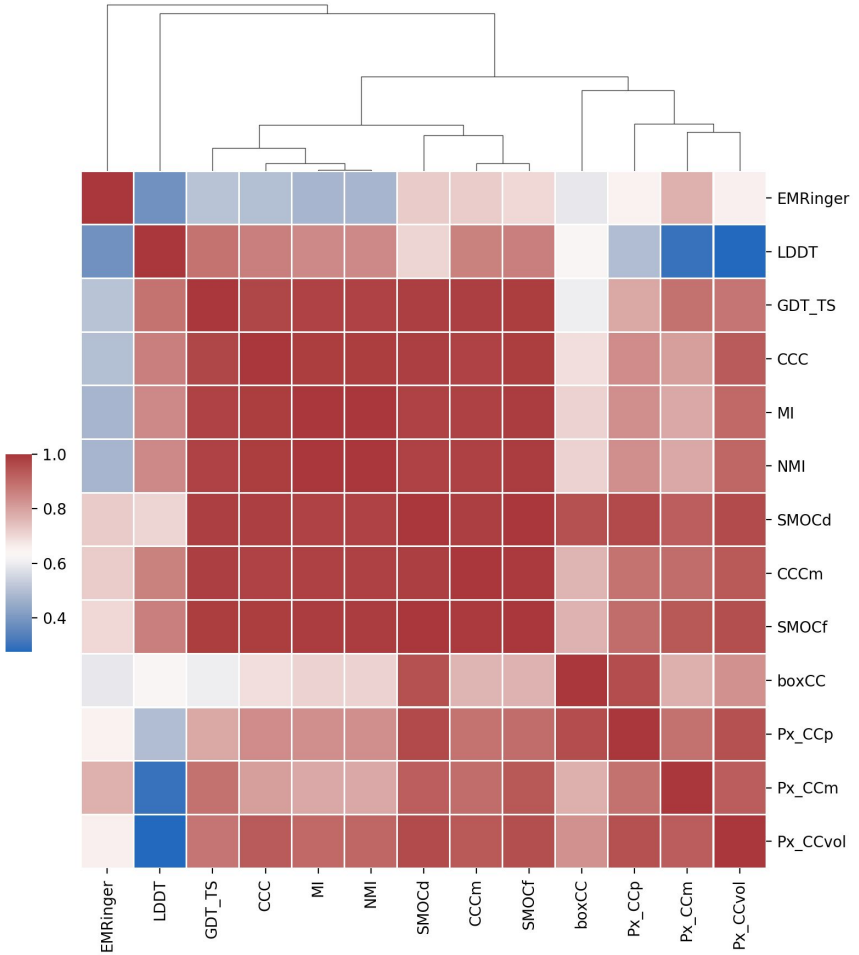
Scoring correlation

For T1094-D2



Scoring correlation

For T1036s1 with clustering



Scores

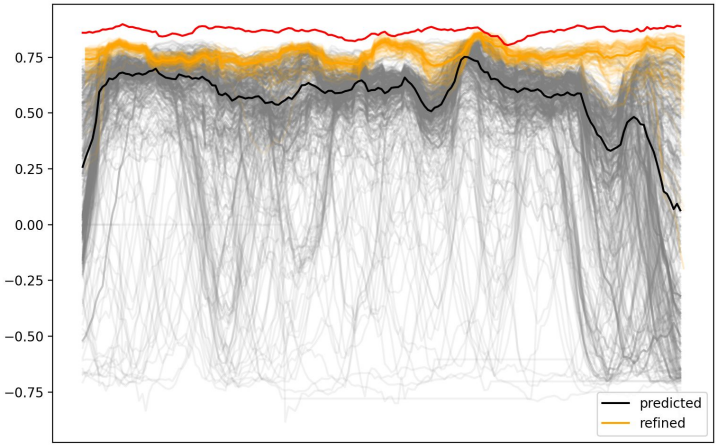
$$CCC = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} * \sqrt{\sum (y - \bar{y})^2}}$$

$$SMOC = \frac{\sum (xy)}{\sqrt{\sum (x)^2} * \sqrt{\sum (y)^2}}$$

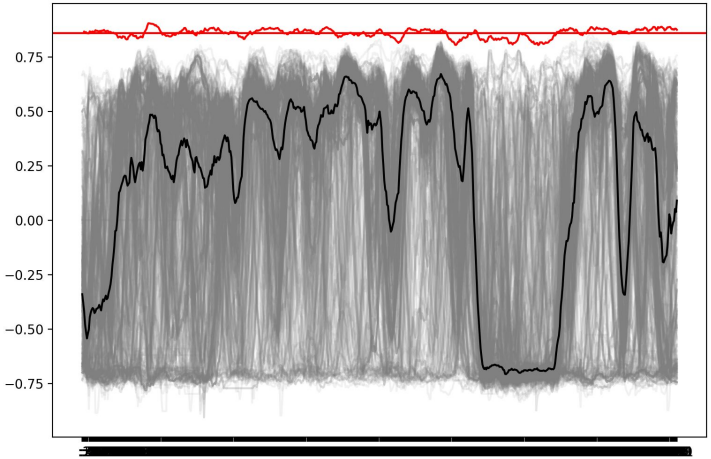
$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Slide with correlations, smocs of all predictions

T1092-D2



T1093



Slide with correlations, smocs of all predictions