Average maximal F1 score on the OOD dataset