Multimodal CCS Update

2021/02/18

Definition of the Different Feature Sets Tested

- **MQN** — predictions made using a model trained on MQNs only
- **MD3D** — predictions made using a model trained on MD3Ds only (PMIs + radial mass distributions)
- **COMB** — predictions made using a model trained on all MQNs + all MD3Ds
- **MIN** — predictions made using a minimal feature set (a few MQNs + PMIs) determined from feature selection trials

Fluoroquinolones (protomers) — Results from Different Feature Sets

** there were 32 plots in total so I went through them and summarized them qualitatively instead of showing all of them*

magnitude relative to measured values

	MQN	MD3D	COMB	MIN
CIPR	I	LL	HH	L
ENOX	I	L	HH	L
ENRO	HI	HI	HH	HI
LEVO	HI	L	HH	L
LOME	HI	L	HH	L
NORF	LI	L	HH	L
ORBI	HH	L	HI	LI
PEFL	HH	HI	HH	HI

low → intermediate → high

- *This table reflects a qualitative measure of generally how close the predicted CCS values are to the measured values*
- *LL - below the lower measured CCS, L - close to the lower measured CCS, LI - intermediate but closer to the lower measured CCS, I - intermediate between measured values, HI - intermediate but closer to higher measured CCS, H - close to higher measured CCS, HH - above the higher measured CCS*
- *The MQN model was probably on average the closest, most often producing intermediate predictions slightly closer to the higher measured CCS value*
- *the MD3D model almost always produced predictions closer to the lower CCS value*
- *the combined model pretty much always significantly over predicted CCS values*
- *the minimal feature set model almost always produced predictions closer to the lower CCS value*

protomer rank order CCS

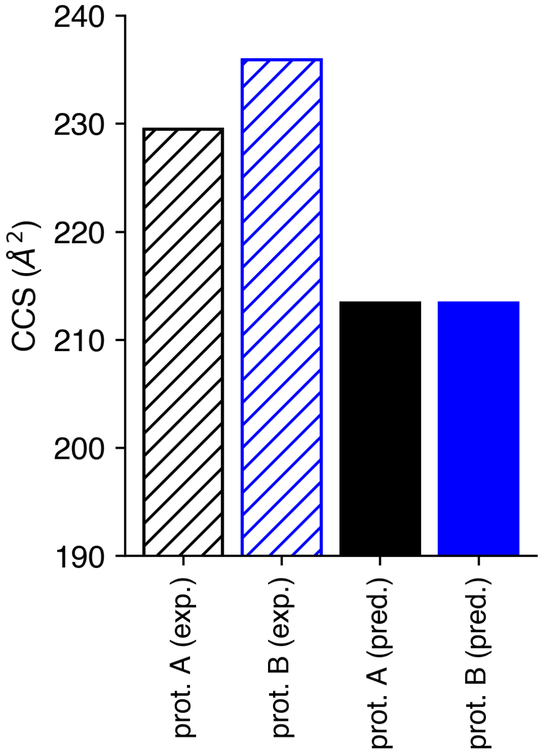
	MQN	MD3D	COMB	MIN
CIPR	C > ABD	C > BD > A	ABCD	C > ABD
ENOX	C > ABD	ACD > B	ABCD	C > ABD
ENRO	C > ABD	ABCD	ABCD	C > ABD
LEVO	C > ABD	C > ABD	ABCD	A > BCD
LOME	AC > BD	ACD > B	C > AD > B	C > B > D > A
NORF	C > ABD	BCD > A	ABCD	C > ABD
ORBI	C > ABD	AC > BD	ABCD	A > C > D > B
PEFL	C > ABD	C > BD > A	ABCD	C > A > D > B

no difference → small difference → large difference

- *This table reflects the rough rank-order of predicted CCS values for the protomers, with the darkness indicating the qualitative size of the differences*
- *The MQN model almost universally predicted pretty small differences between the protomers but interestingly with the C protomer (central ring nitrogen) predicted slightly larger — I expect this is attributable to a topology change in this protomer relative to the others since their compositions are the same*
- *MD3D produced some larger differences, albeit with some more mixed results — we do often see the C protomer larger than others though*
- *the combined model almost always produced indistinguishable predictions for all protomers — it probably learned to recognize a bunch of extraneous features that do not differ much between these protomers*
- *The MIN model produced medium to weak differences but also mostly predicted the C protomer to be larger than the others*

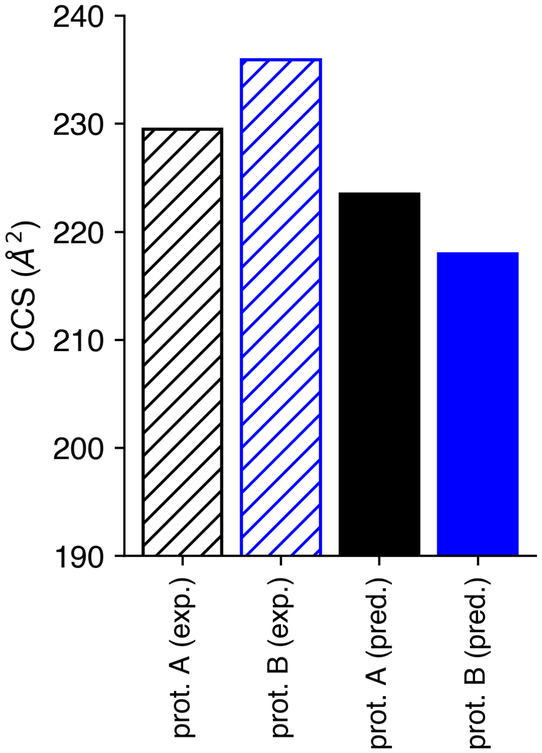
Cefpodoxime Proxetil (protomers) — Results from Different Feature Sets

MQN



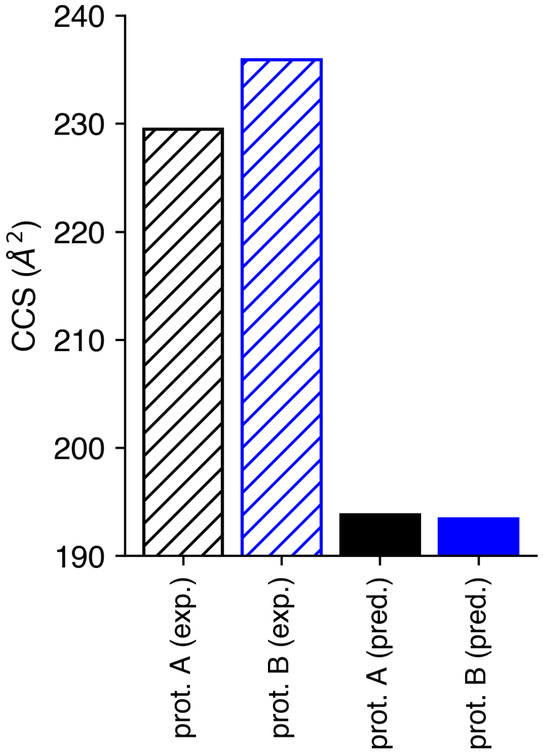
- *no difference between the predicted values — no topological or compositional difference between these protomers so same MQNs*
- *both values under predicted*

MD3D



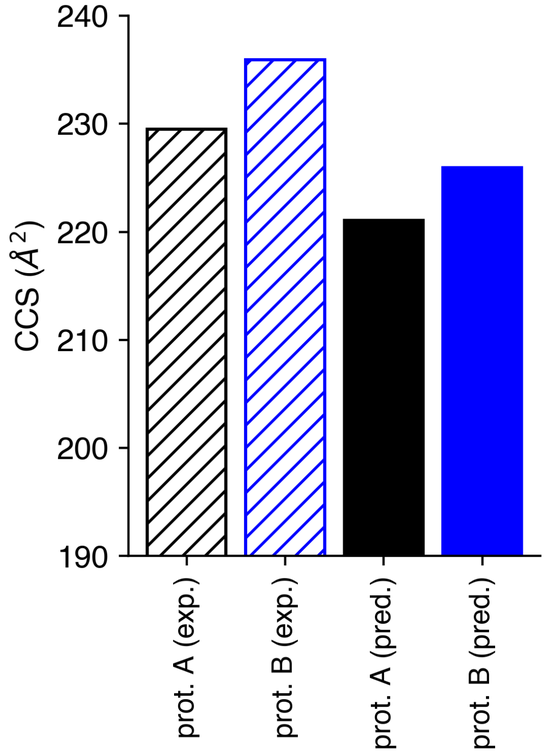
- *both values under predicted, opposite trend to previous results*
- *Something about these protomer structures is not captured well by the MD3Ds alone*

COMB



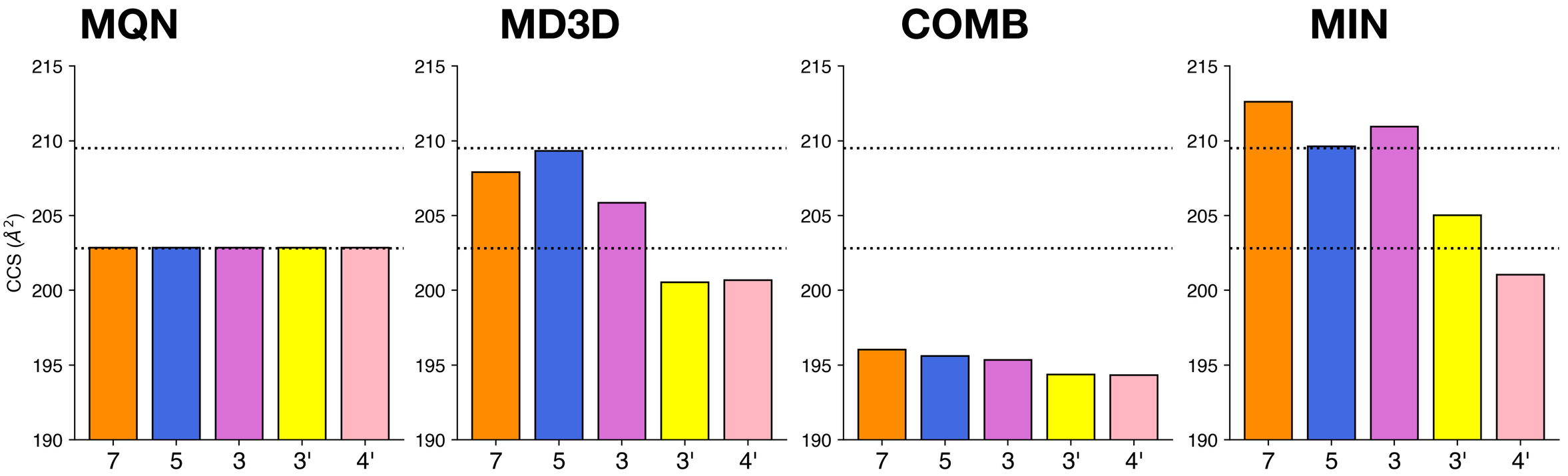
- *both values under predicted, very slight similar trend to MD3D*
- *the combined model probably learned to recognize a bunch of extraneous features that do not differ much between these protomers*

MIN



- *both values under predicted, same trend as previous results*
- *the combination of a few MQNs and the PMIs seem to capture the differences between these protomers correctly*

Quercetin Glucuronides (positional isomers) — Results from Different Feature Sets



- *no difference between isomers by MQNs — as expected*
- *Interestingly they all match the lower CCS, so you could possibly say that for this case some of the positional isomers present larger than you would expect given its composition?*

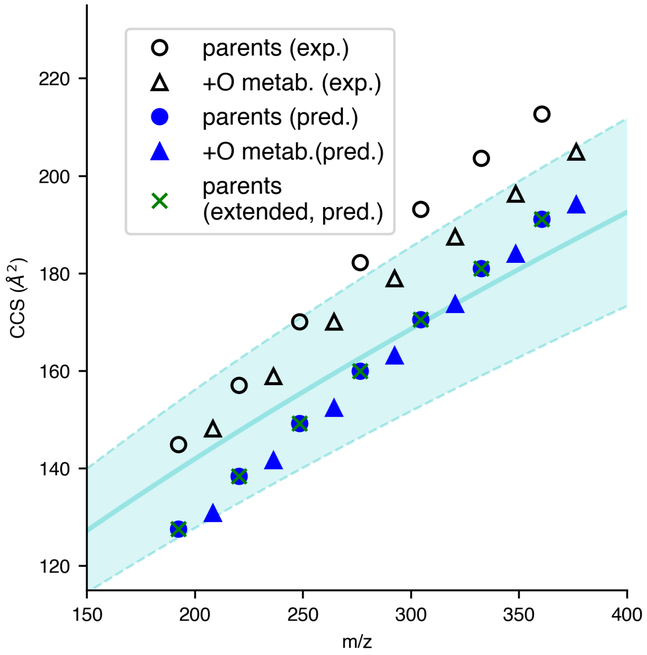
- *MD3D model captures positional isomer differences pretty well, only discrepancy is the 3-isomer which was previously reported as closer to the lower CCS*
- *slight systematic under prediction*

- *All values under predicted, very slight similar trend to MD3D*
- *the combined model probably learned to recognize a bunch of extraneous features that do not differ much between these isomers*

- *same trend as MD3D, but closer to measured values*
- *a difference also emerges from the 3’ and 4’ isomers, probably from different weighting of the PMIs in this feature set*

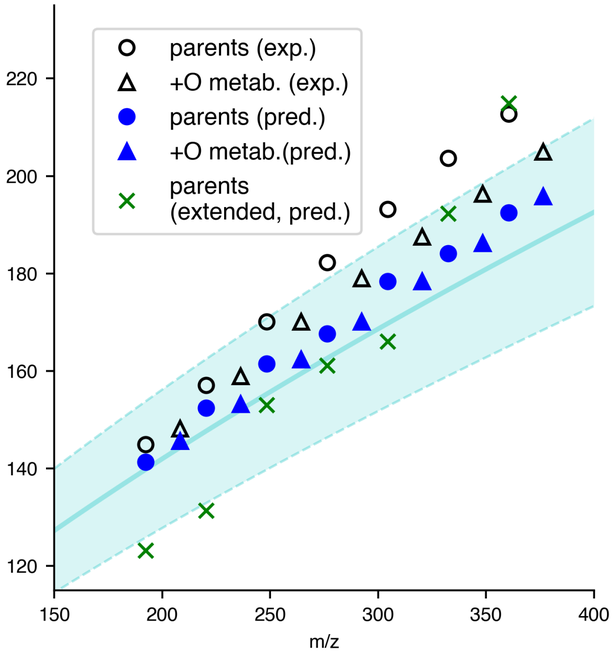
BACs (compaction in +O metabolites) — Results from Different Feature Sets

MQN



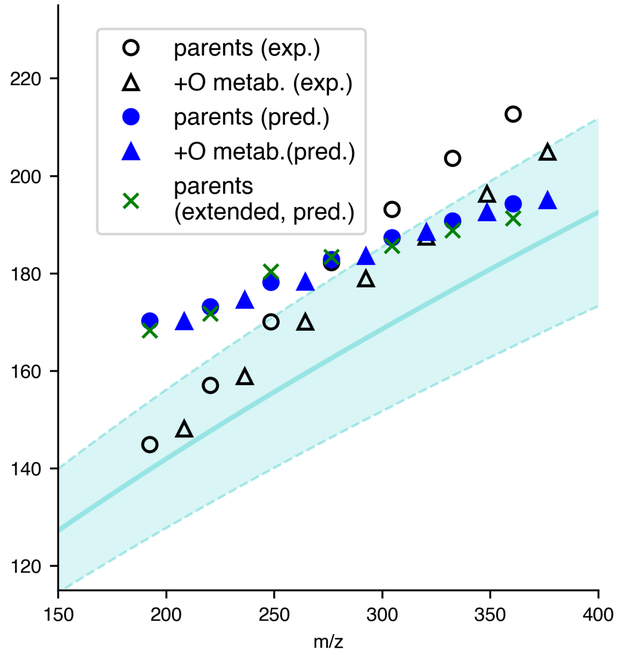
- *no difference between fully extended and parent structures — expected because MQNs do not capture conformation*
- *parents and metabolites present with uniform CCS-m/z relationships*
- *interestingly the metabolites occupy a slightly more dense trend, I suspect this is due to increasing the heavy atom count but doing so with a more dense (O) atom — also slightly different topology (different counts of nodes/ edges/connectivity for -CH3 vs. -OH)*
- *overall, magnitude of predicted CCS is lower than measured*

MD3D



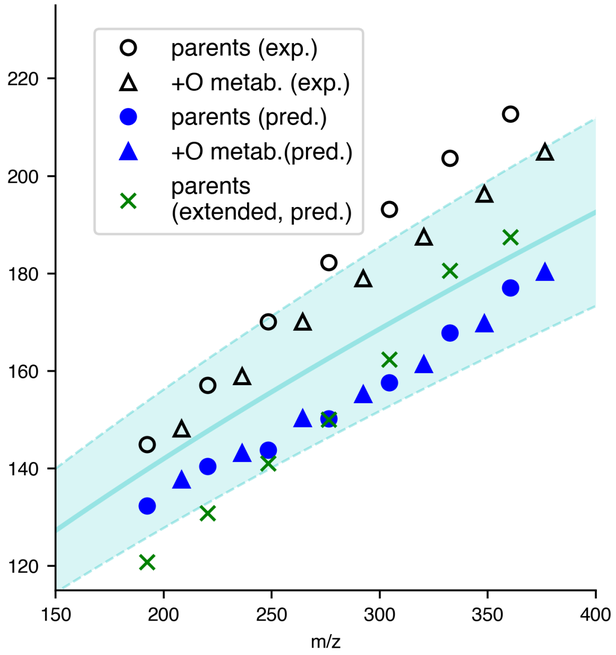
- *This feature set produces the closest predictions to the experimental data overall — in contrast to MQNs this is likely from lack of representation of lipid like structures in training data, the MD3Ds are probably a bit more generalizable to different chemical classes than MQNs just based on how they are computed*
- *The short chain values are pretty close, with increasing length the predicted values are systematically under predicted and parents fail to separate from metabolites*
- *The extended conformer predictions are weird, for short chains they are lower than the ensemble of modeled structures and for long chains they are higher. Maybe a consequence of their having large PMI1 and small PMI2 and PMI3 in addition to the steady increase in the proportions of higher distances in the RMD?*

COMB



- *Not a lot of response to different conformations, chain length, or composition in CCS for the predicted values (pretty flat slope) — the combined feature set model probably responds to a bunch of extraneous features that do not really differ much between this group of compounds*

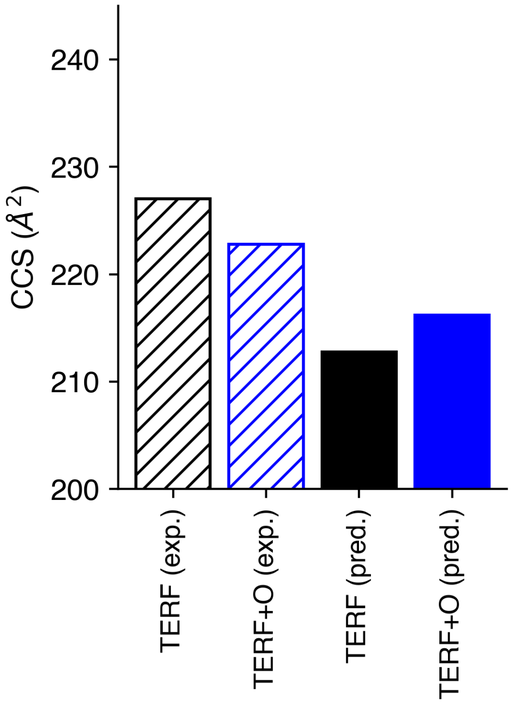
MIN



- *minimal feature set model similar to MD3D model but with slightly smaller effect size — can probably be attributed to the inclusion of some MQNs*
- *MIN only has the PMIs from MD3D and shows the same trend for the extended conformers, so probably the PMIs are what is driving the interesting behavior*

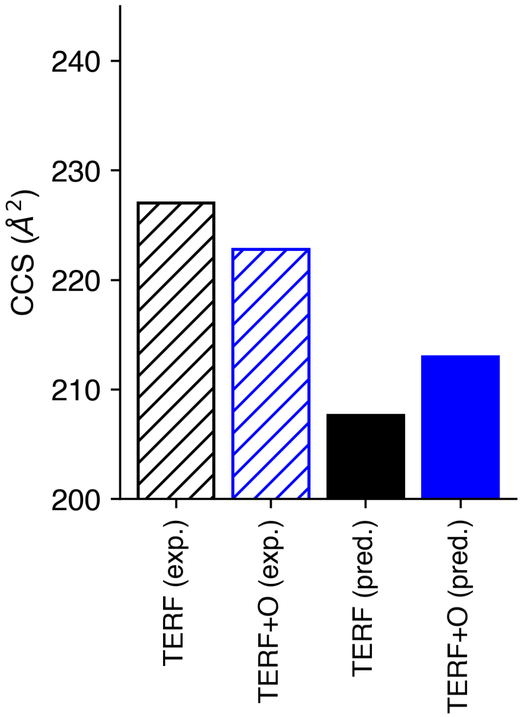
Terfenadine (compaction in +O metabolites) — Results from Different Feature Sets

MQN



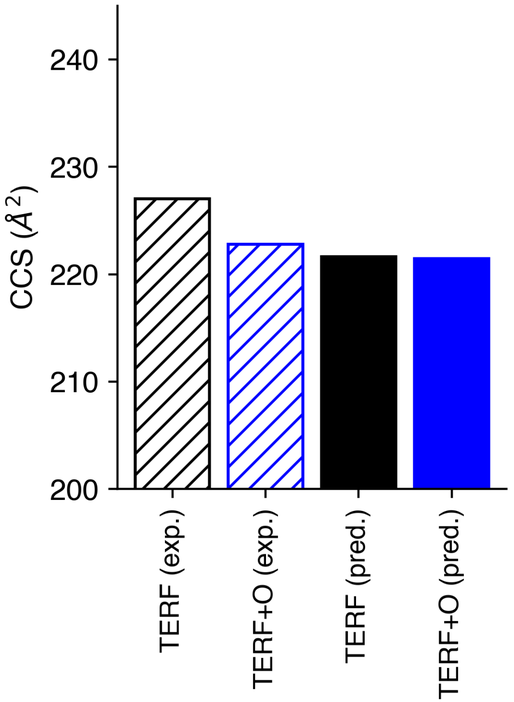
- *both values under predicted*
- *like with the BACs the MQNs seem to produce the "expected" increase in CCS corresponding to addition of O*

MD3D



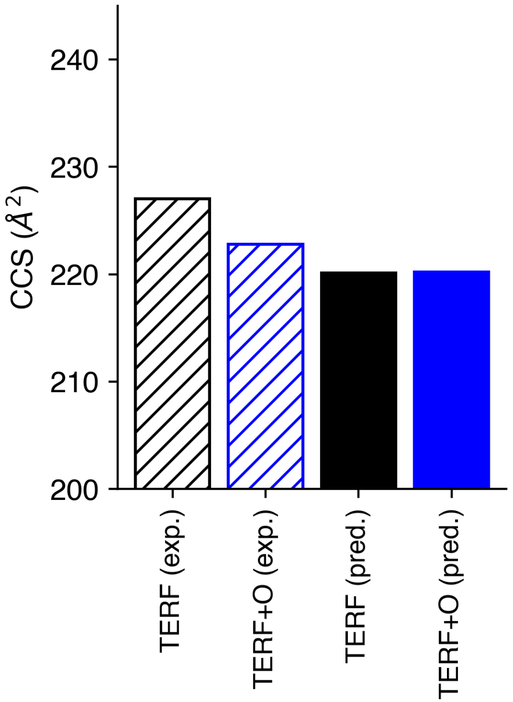
- *both values under predicted*
- *with the BACs the MD3D model capture a small degree of compaction in some of the intermediate chain BACs, not the case with TERF*

COMB



- *both values under predicted, but less than MQN or MD3D*
- *not much difference between parent and metabolite, but interestingly there is a *very slight* compaction in the metabolite*

MIN



- *both values under predicted, but less than MQN or MD3D*
- *no real difference between parent and metabolite*

Summary

- It seems that in some cases (protomers, positional isomers) observed trends in CCS can be recapitulated using models trained on one or more of the feature sets I tested, in general I think the MIN set seems to do the best overall

- For compaction in +O metabolites I don’t think there is enough representation in the training data to correctly predict it. Further I just don’t think the BACs (being very lipid-like, esp. longer chain BACs) are really well represented in the chemical space covered by the training data.

- In essentially all cases, the COMB model performs poorly. I think this can be attributed to overfitting from all the extraneous variables in the feature set, this comports well with the bulk performance metrics (very low training set error, much higher test set error) which indicate poor generalizability