



UNIVERSITÄT
LEIPZIG



DFG Deutsche
Forschungsgemeinschaft
German Research Foundation

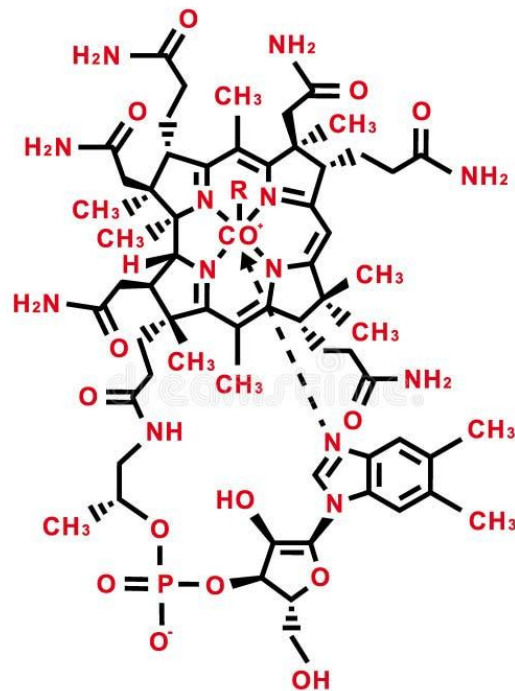
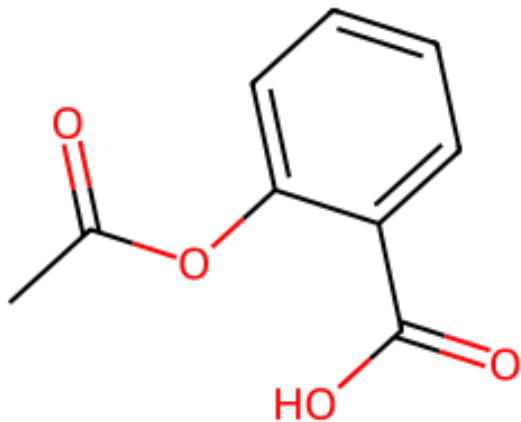
41st TBI Winterseminar in Bled

ENCODING CHEMICAL INTUITION: THE CHALLENGE OF SYNTHESIZABILITY PREDICTION

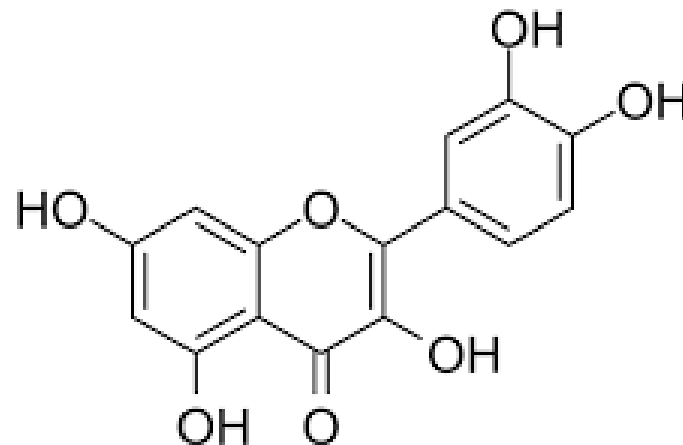
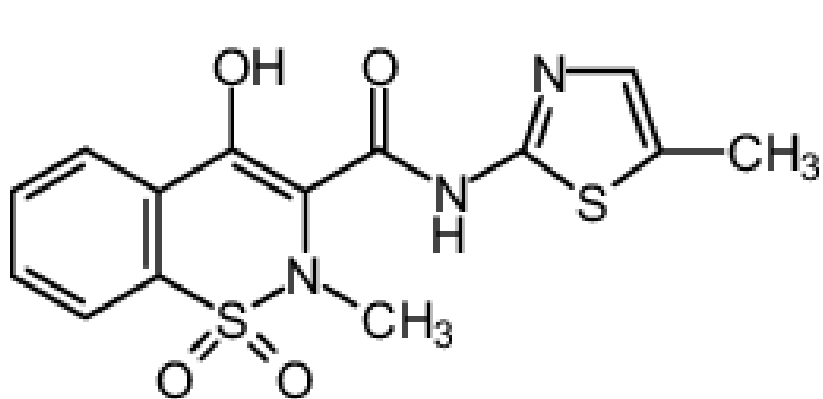
Arina Shelashen, PhD

Bled, 11 Feb 2026

CAN WE PREDICT SYNTHESIZABILITY OF THE MOLECULE?



CAN WE PREDICT SYNTHESIZABILITY OF THE MOLECULE?



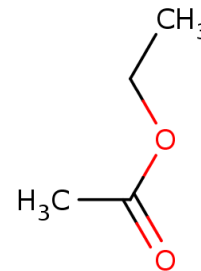
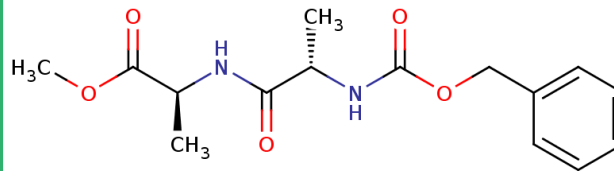
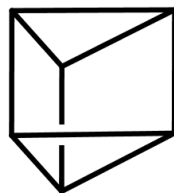
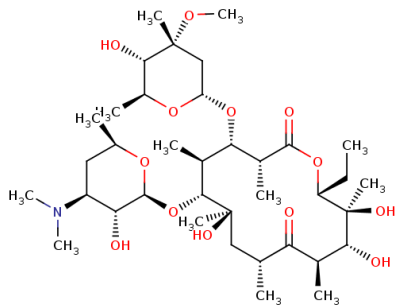
SYNTHESIZABILITY SCORES

	SA	Training data
SA	Structure based	PubChem (10 ⁶ molecules)
SC	Reaction based	Reaxys
RA	Reaction based	outcomes generated by AiZynthFinder (USPTO based) on ChEMBL compounds
GASA	Reaction based	outcomes generated by Retro* on ChEMBL compounds, ZINC, Nonpher
DeepSA	Structure based	outcomes generated by Retro* on ChEMBL compounds, ZINC, Nonpher
DFR	Structure based	PubChem + labels via FRA
MolPrice	Structure based	Molport (purchasable database)

TEST DATASET

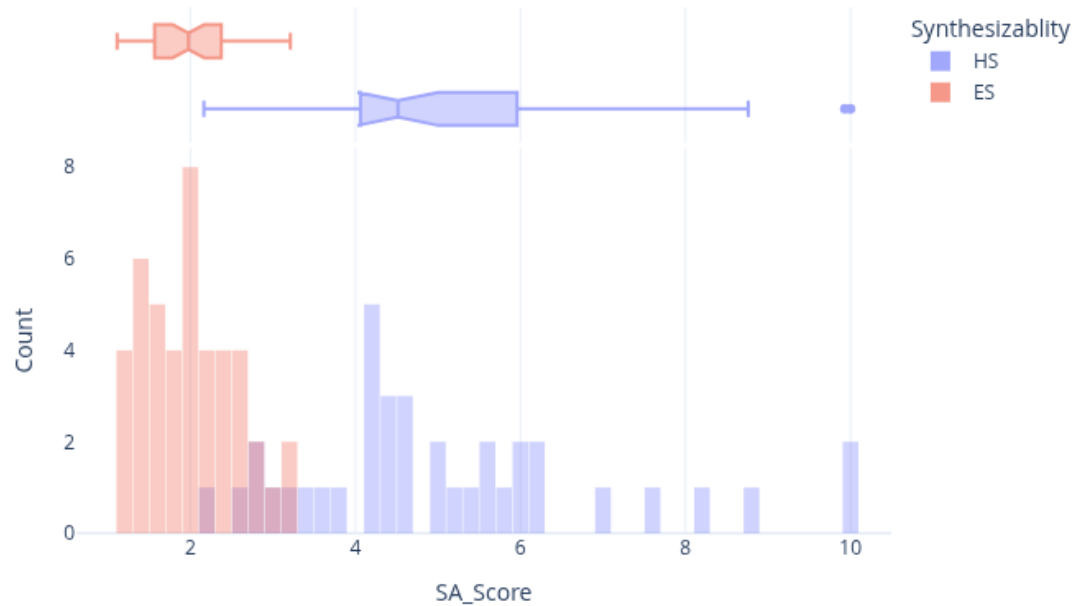
81 compounds

37 HS / 44 ES

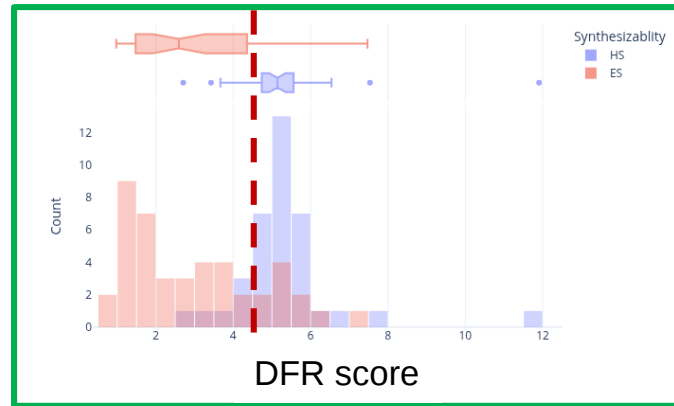
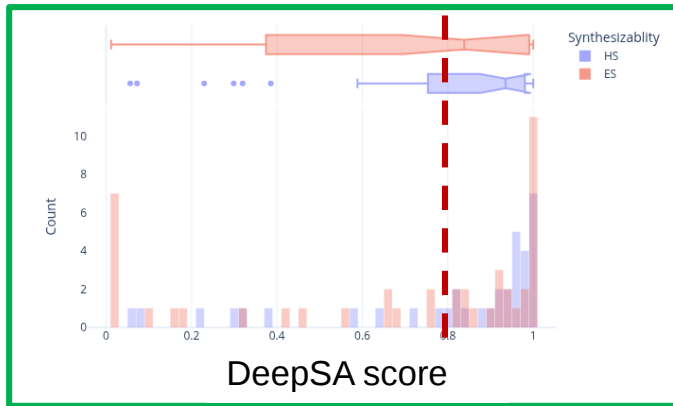
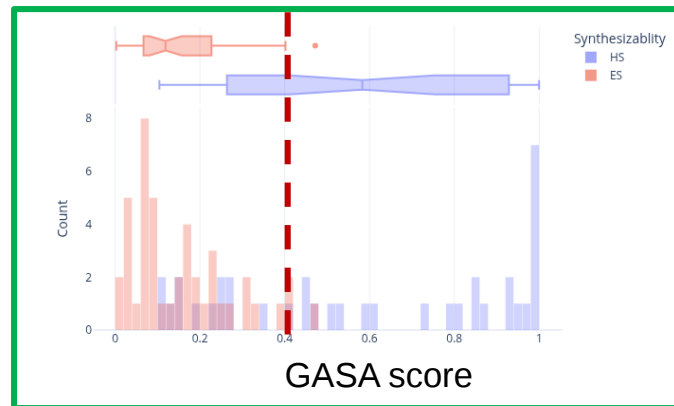
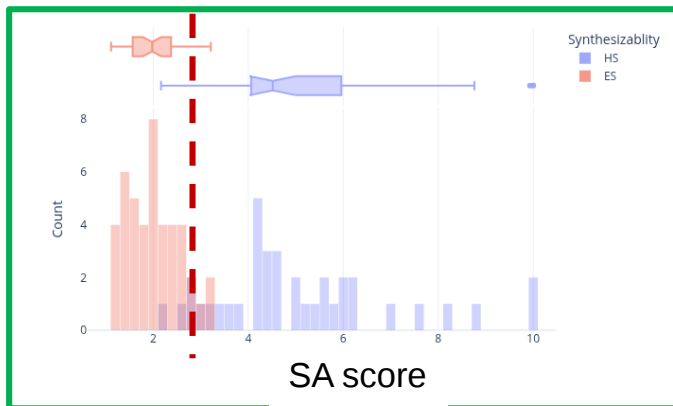


TEST DATASET

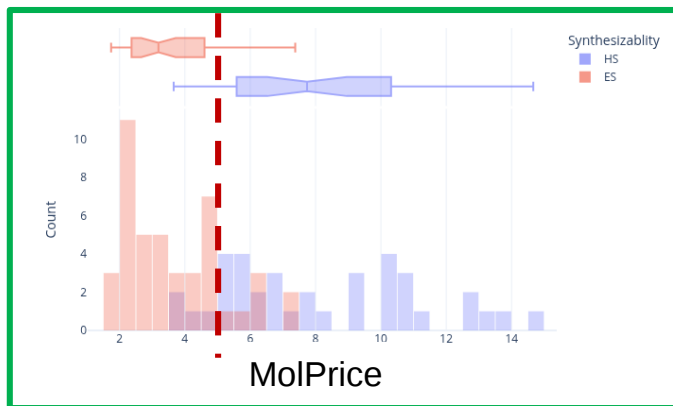
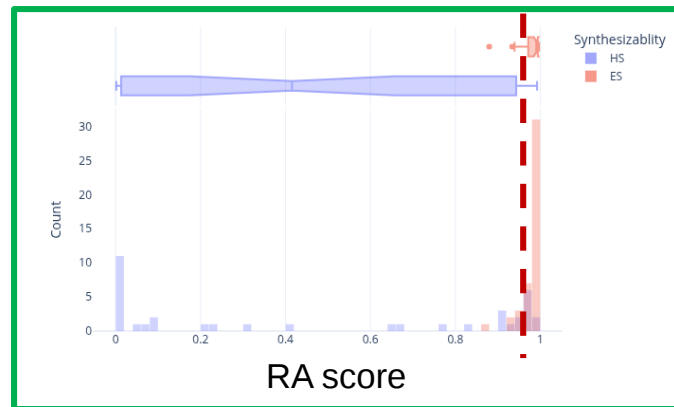
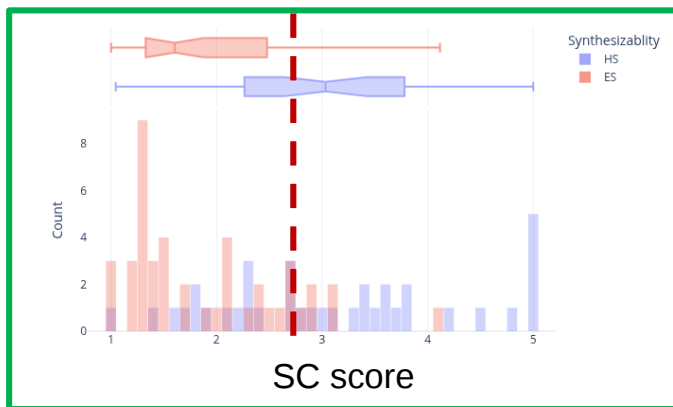
Distribution of SA_Score by Synthesizability



TEST DATASET

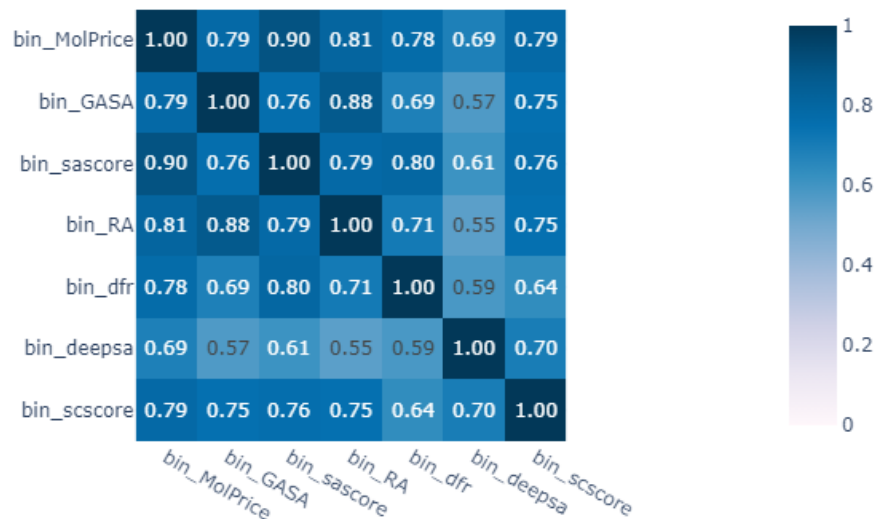


TEST DATASET



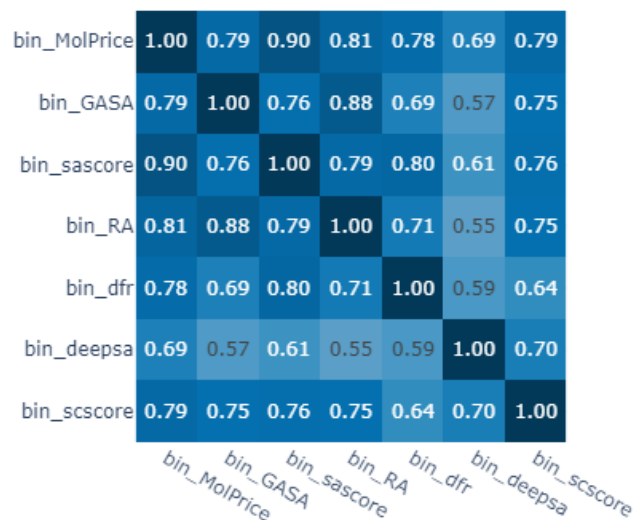
TEST DATASET

Pairwise Agreement Between Binary Classifiers



TEST DATASET

Pairwise Agreement Between Binary Classifiers

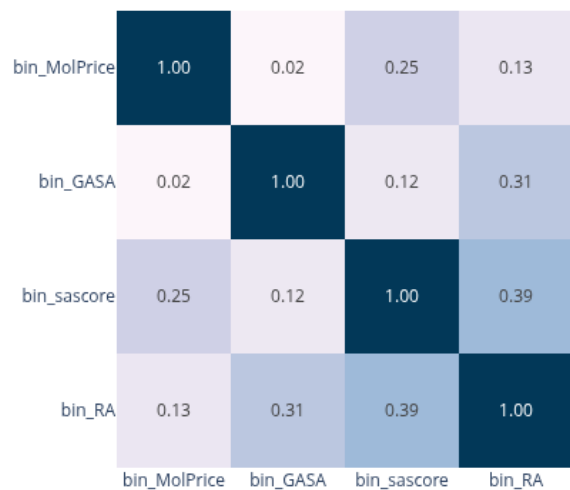


Cohen's Kappa Between Tools



USPTO / ZINC DATASETS

Cohen's Kappa Between Tools



Cohen's Kappa Between Tools



LOGISTIC REGRESSION MODEL

USE SEVERAL SCORES SIMULTANEOUSLY → LOGISTIC REGRESSION

LOGISTIC REGRESSION MODEL

USE SEVERAL SCORES SIMULTANEOUSLY → LOGISTIC REGRESSION

EXHAUSTIVE SEARCH BY F1 SCORE

$$z = -2.11 SA_{score} - 0.93 GASA_{score} + 0.35 RA_{score} + 6.37$$

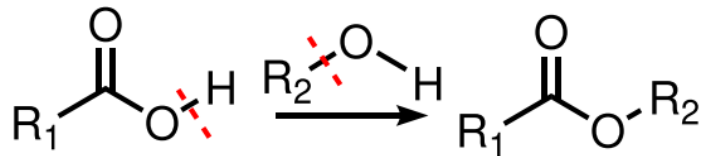
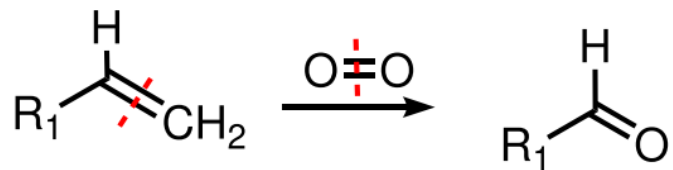
$$P(ES) = \frac{1}{1 + e^{-z}}$$

PERFORMANCE COMPARISON

Predictor	Accuracy	Precision	Recall	F1
Predicted_Synthesizability	0.925	0.880	1.000	0.936
bin_sascore	0.912	0.951	0.886	0.918
bin_RA	0.850	0.796	0.977	0.878
bin_MolPrice	0.863	0.902	0.841	0.871
bin_GASA	0.825	0.788	0.932	0.854
bin_DFR	0.813	0.914	0.727	0.810
bin_scscore	0.725	0.720	0.818	0.766
bin_deepsa	0.575	0.679	0.432	0.528

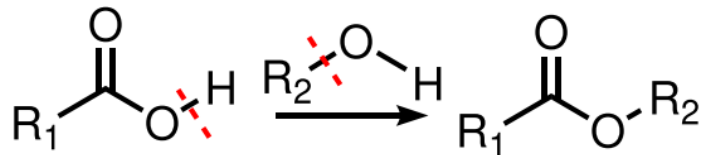
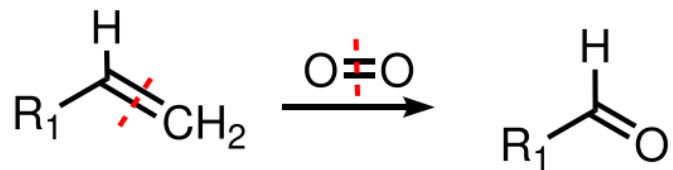
APPLICATION ON CUT & JOIN RESULTS

C&J idea



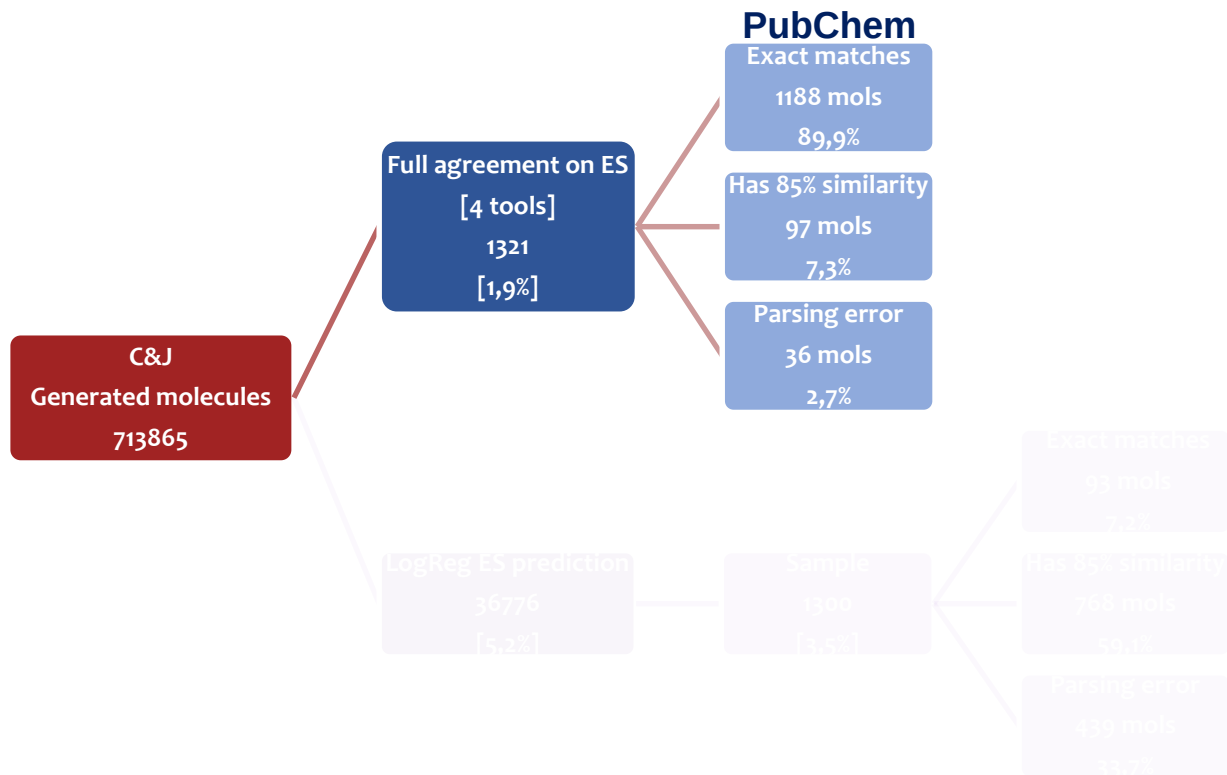
APPLICATION ON CUT & JOIN RESULTS

C&J idea

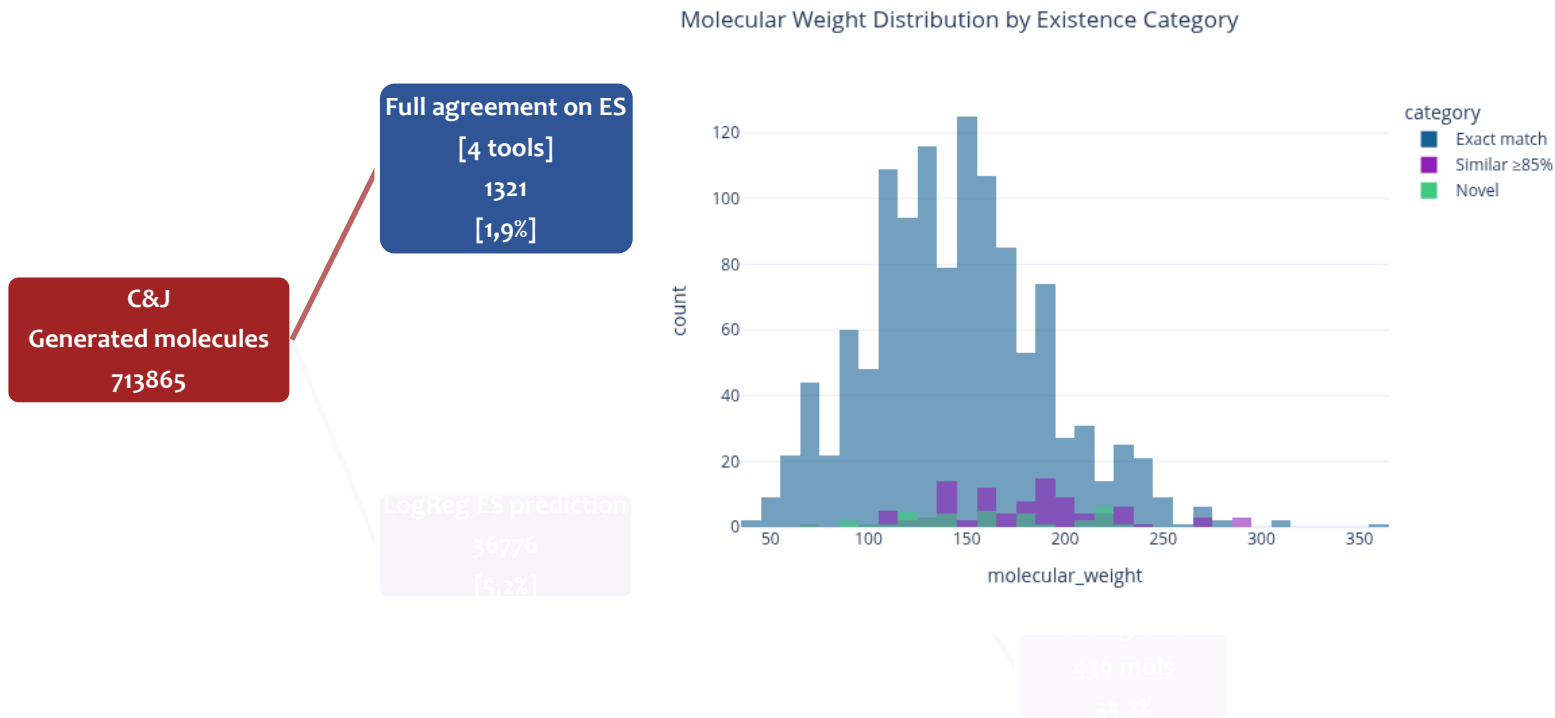


3 input molecules $\rightarrow$ 150.000 generated molecules

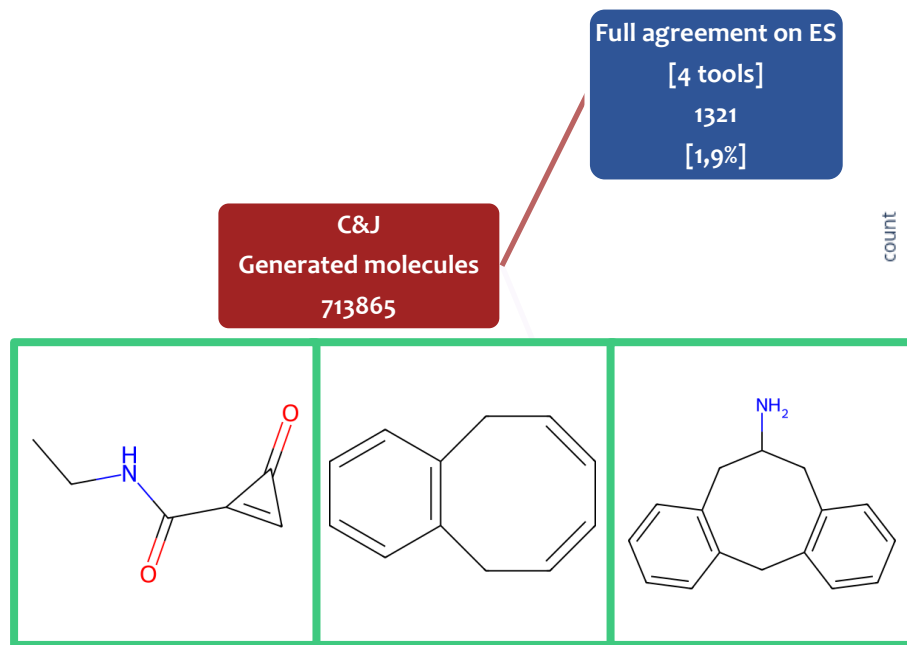
APPLICATION ON CUT & JOIN RESULTS



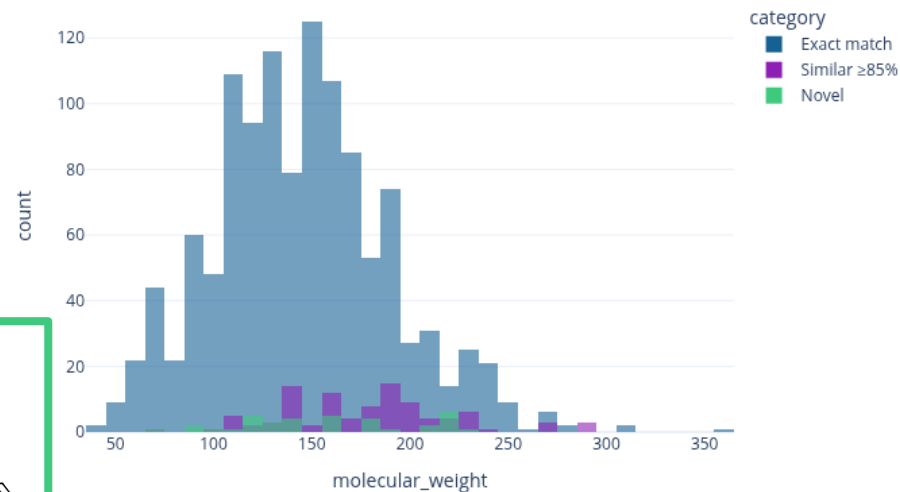
APPLICATION ON CUT & JOIN RESULTS



APPLICATION ON CUT & JOIN RESULTS



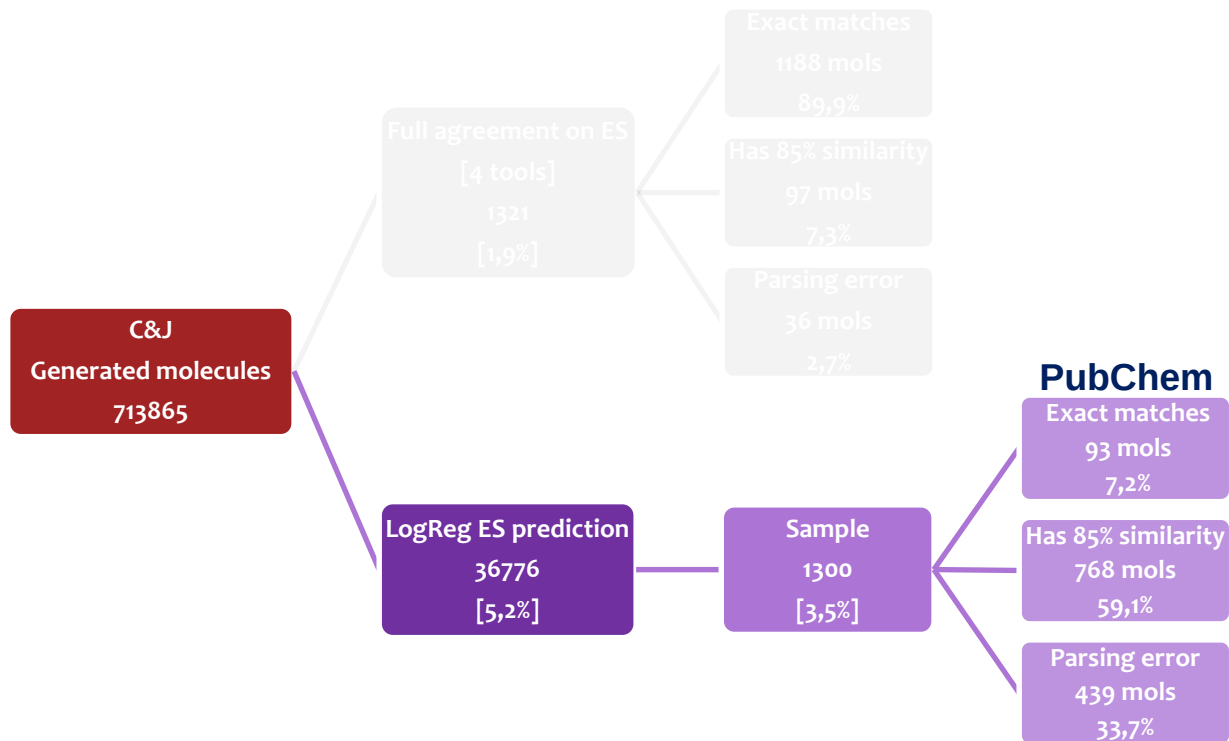
Molecular Weight Distribution by Existence Category



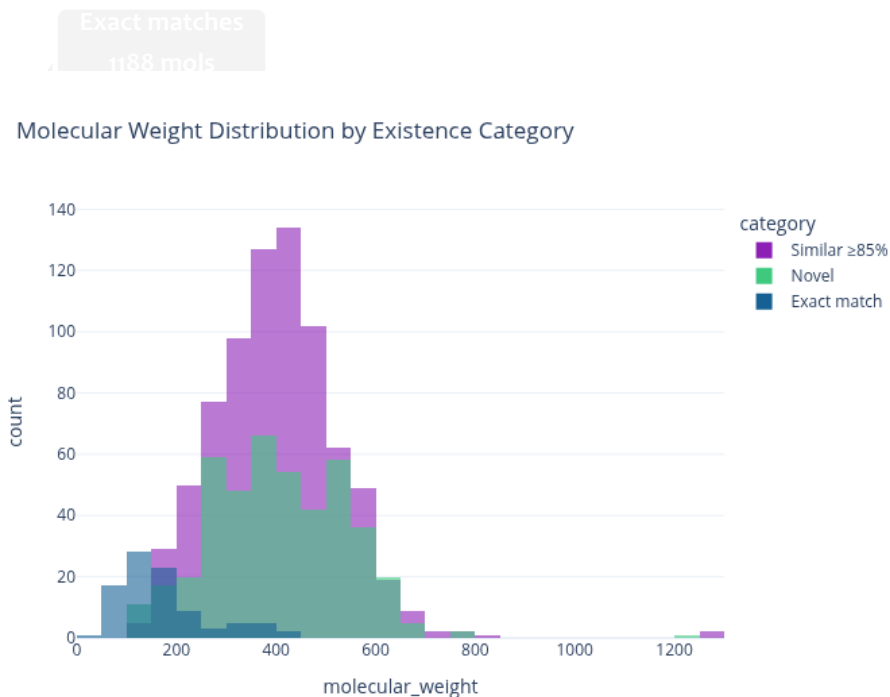
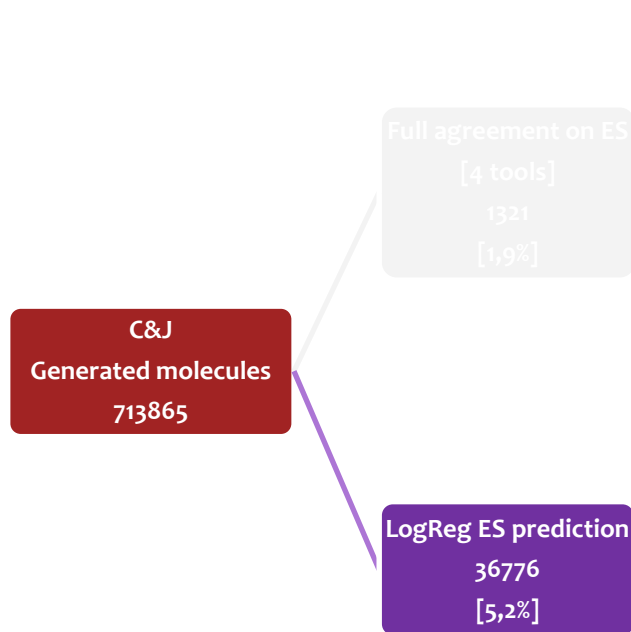
439 mols

0.6%

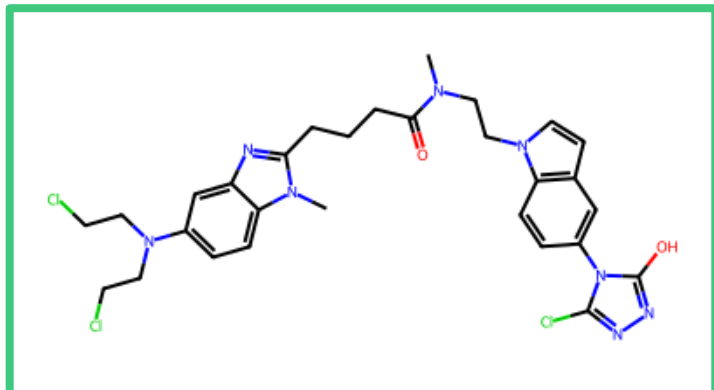
APPLICATION ON CUT & JOIN RESULTS



APPLICATION ON CUT & JOIN RESULTS



APPLICATION ON CUT & JOIN RESULTS

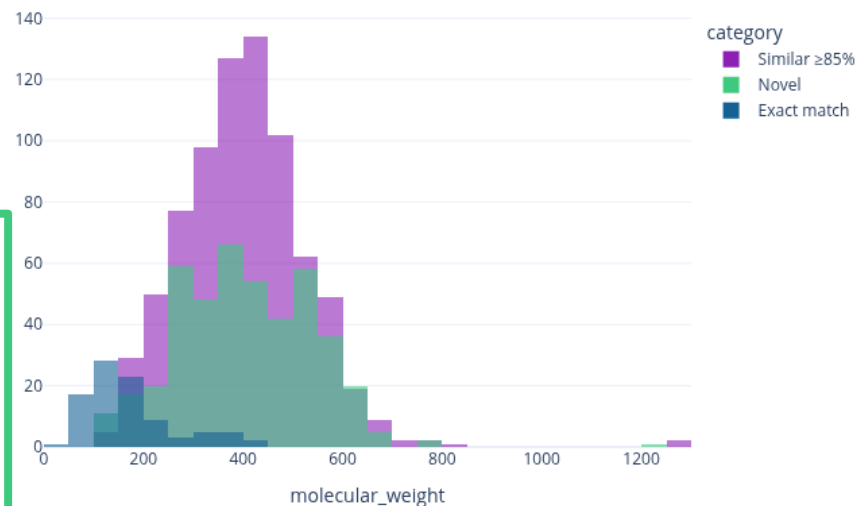


h ES

Exact matches

1188 mols

Molecular Weight Distribution by Existence Category



CONCLUSION

- ❖ Synthesizability prediction remains challenging
- ❖ Existing scores help eliminate unrealistic candidates



UNIVERSITÄT
LEIPZIG

THANK YOU FOR YOUR ATTENTION!

Arina Shelashen

PhD student

Bioinformatics Leipzig

arina@bioinf.uni-leipzig.de

LOGISTIC REGRESSION MODEL

USE SEVERAL COMPOUNDS SIMULTANEOUSLY → LOGISTIC REGRESSION

EXHAUSTIVE SEARCH BY F1 SCORE

$$z = -2.11 SA_{score} - 0.93 GASA_{score} + 0.35 RA_{score} + 6.37$$

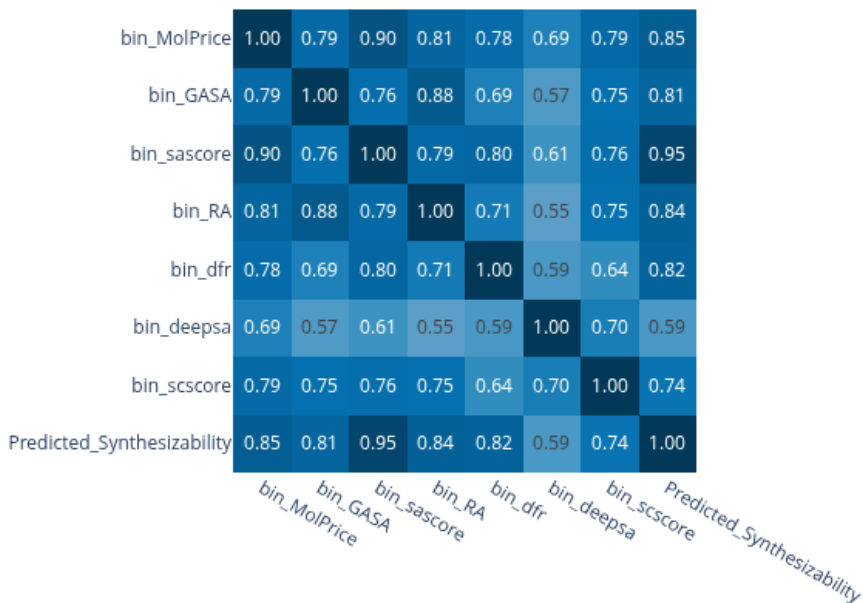
$$P(HS) = \frac{1}{1 + e^{-z}}$$

Scaled:

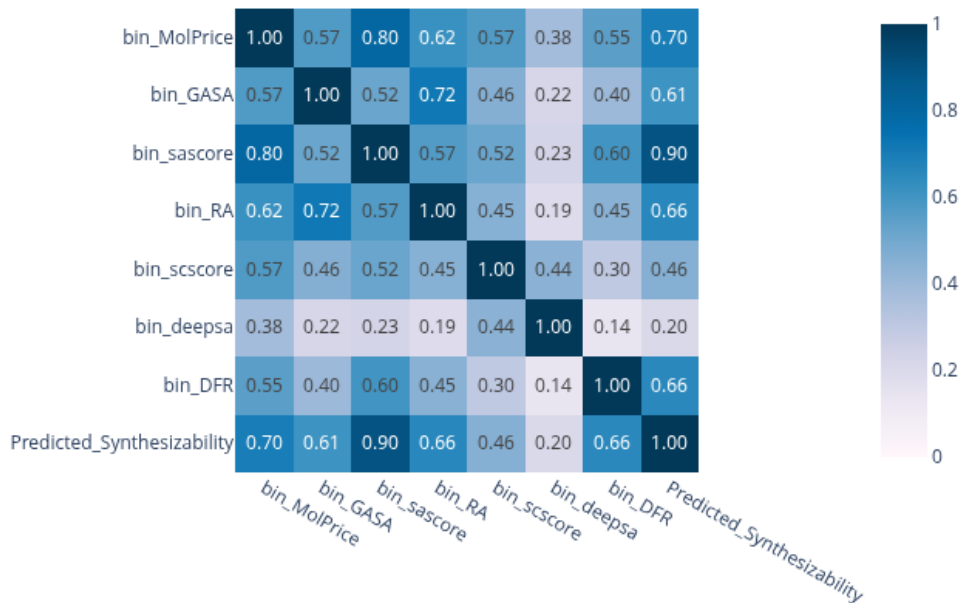
$$z = -2.48 SA_{score} - 1.26 GASA_{score} + 0.64 RA_{score} - 0.85$$

TEST DATASET

Pairwise Agreement Between Binary Classifiers

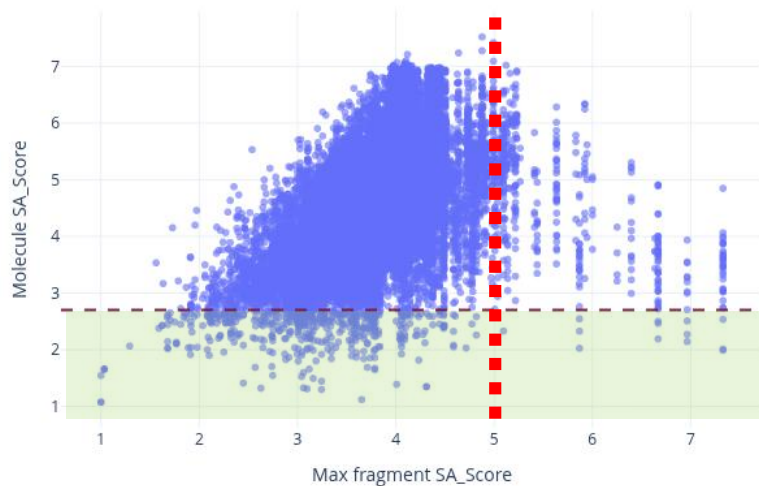


Cohen's Kappa Between Tools

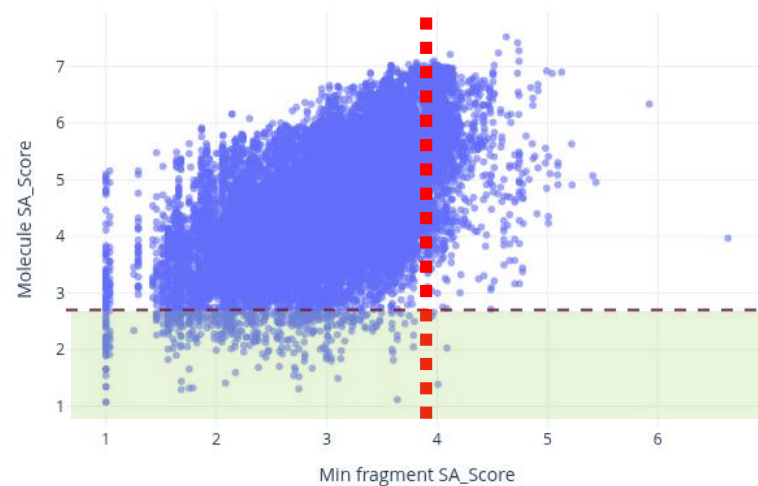


C&J RESULTS

Max fragment score vs molecule score

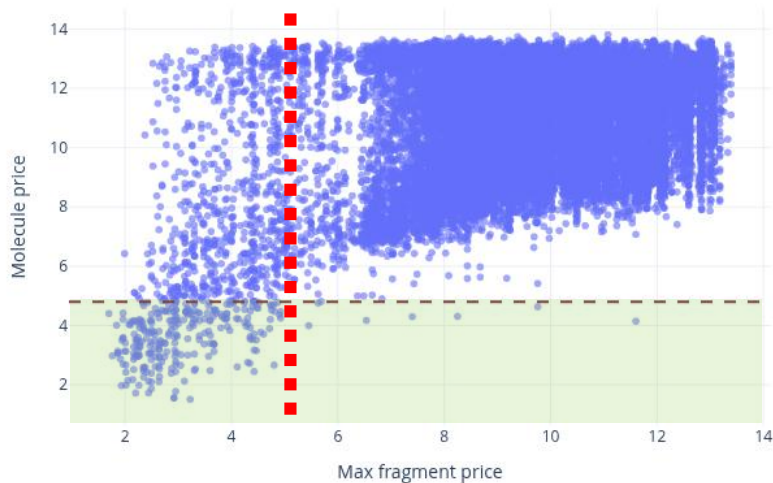


Min fragment score vs molecule score



C&J RESULTS

Max fragment score vs molecule score



Min fragment score vs molecule score

