tmRDKit: 200k RDKit-parseable SMILES strings of transition metal complexes from the CCDC

Jan H. Jensen
Department of Chemistry,
University of Copenhagen

With Maria Rasmussen, Magnus Strandgaard, Julius Seumer,
Angelo Frei (York), David Balcells

@janhjensen

What

SMILES for TMCs with bond orders, formal charges, atom types, etc.

Why?

Synthetic accessibility for *de novo* TMC catalyst design

Test xyz2mol for TMCs

Baseline models (e.g. RF/ECFP4) for ML on TMCs

Options

CCSD has SMILES but many cannot be read by RDKit
w/o `sanitize=False` (no aromaticity or implicit H's)

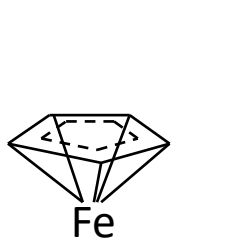
CCSD has coordinates (including H's) but xyz2mol difficult to apply to TMCs

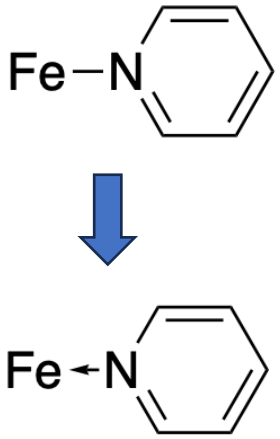
xyz2mol needs the atomic valence and molecular charge
No single atomic valence (oxidation state) for TMs

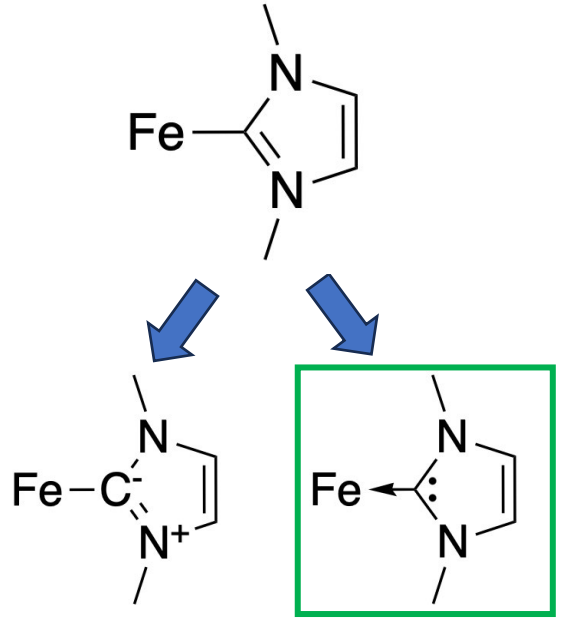
Fixing CCDC SMILES

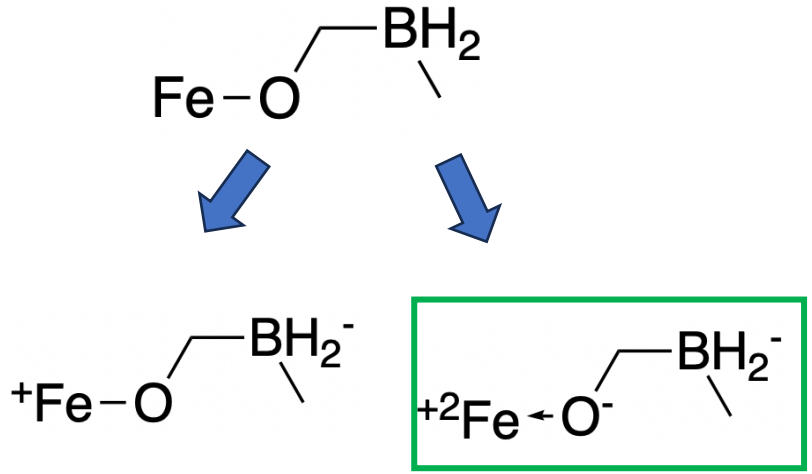
Main problems

`KekulizeException` & `AtomValenceException`

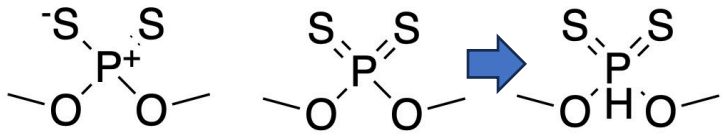


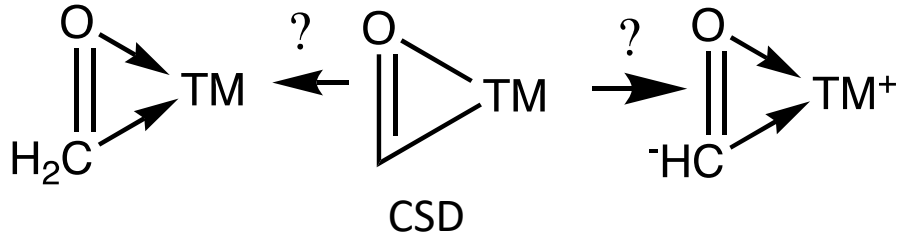






Countless smaller problems
(many not fixed)





60,799 TMCs → 53,320 SMILES

SMILES from xyz2mol (DFT)

Magnus Strandgaard

Deep learning metal complex properties with natural quantum graphs†

Hannes Kneiding, [a] Ruslan Lukin,[a] Lucas Lang, [a] Simen Reine, [a] Thomas Bondo Pedersen, [a] Riccardo De Bin [b] and David Balcells *[a]

2023

tmQMg

Determine ligand charges from DFT NBO calculations on the TMCs

DFT calc. → bond & LP MOs → assign to ligands → **ligand charge** = # electrons - ∑ nuclear charges

→ TM-ligand bonds

TM oxidation state = total charge + ∑ ligand charges

xyz2mol (ligand charges) → ligand SMILES + TM-L bonds → TMC SMILES

60,799 TMCs → 44,737 SMILES

Not all CSD entries have SMILES

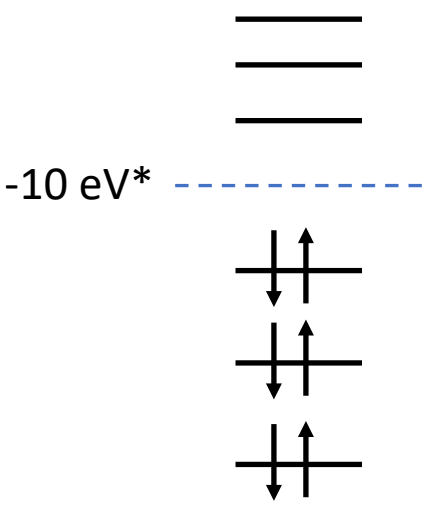
Maria Rasmussen

SMILES from xyz2mol (Hückel)

Individual Hückel calcs on ligands → Fill MOs up to threshold → **ligand charge** = # electrons - ∑ nuclear charges

Connectivity from distances (a la OpenBabel)

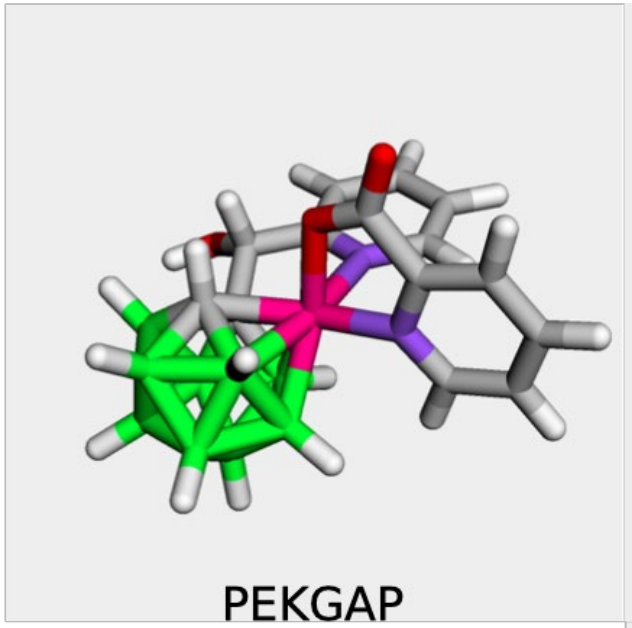
Hückel MOs



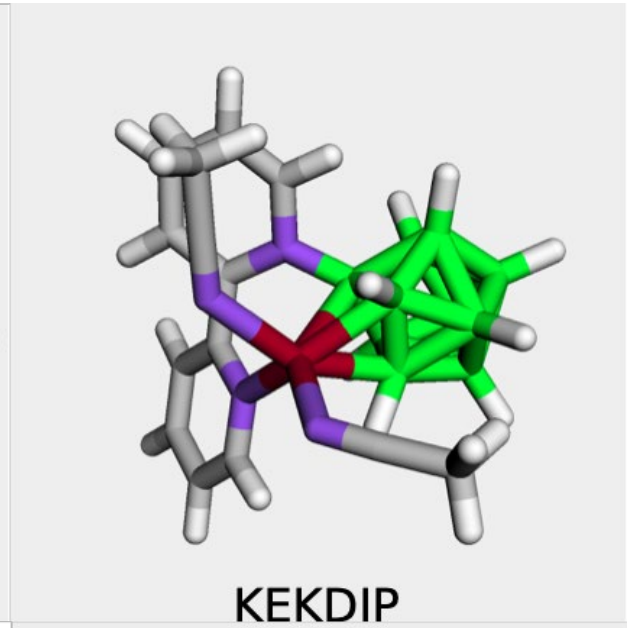
*-9 eV if -10 eV gives positive ligand;
-10.2 eV if -10 eV gives -2 ligand charge

60,799 TMCs → 59,886 SMILES

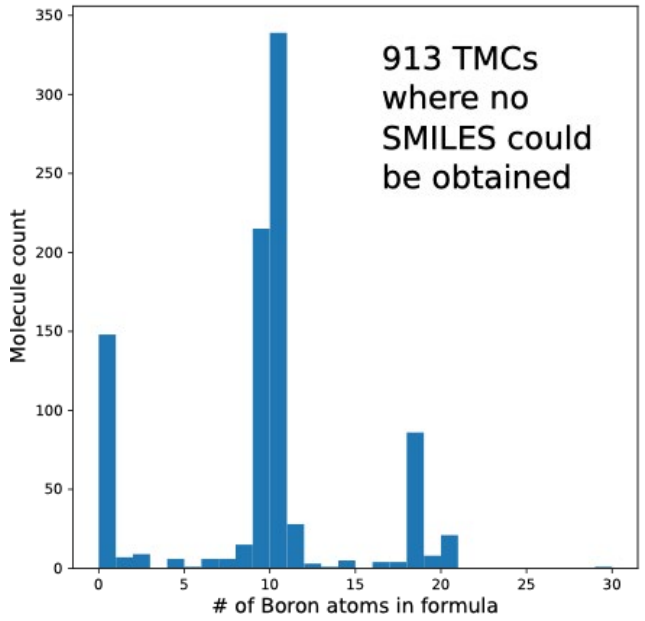
Hückel approach: most fail because of Boron

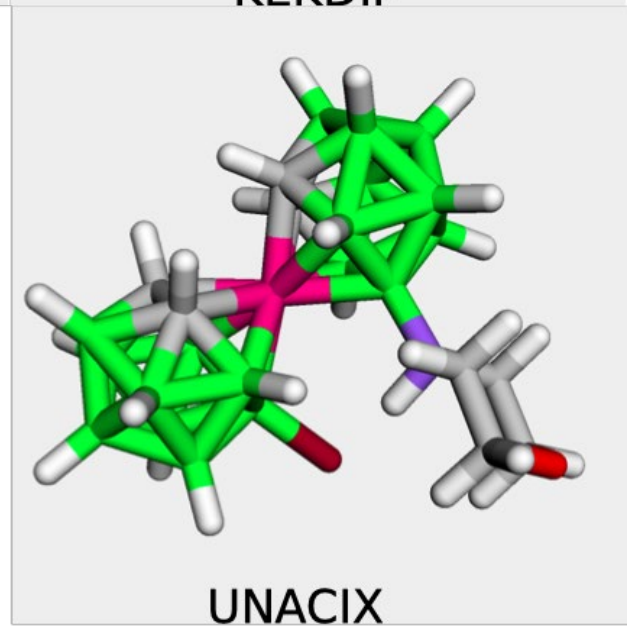


PEKGAP



KEKDIP



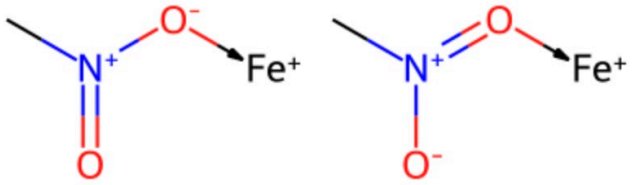


UNACIX

Comparing SMILES

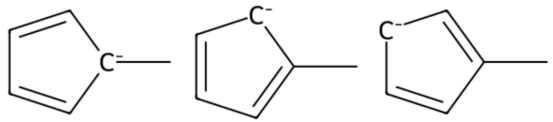
ResonanceMolSupplier

Same TMC SMILES?

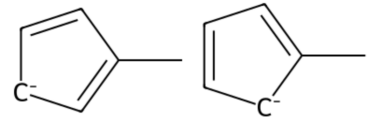


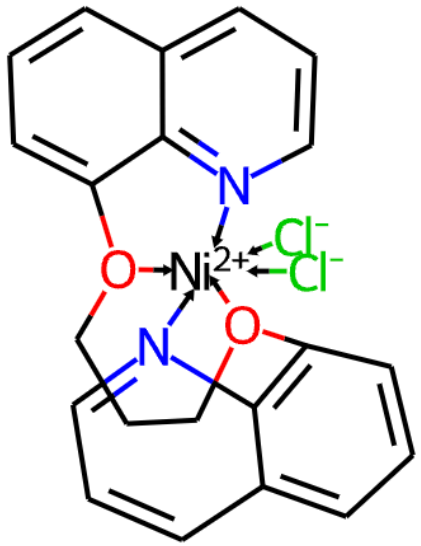
MetalDisconnector

Same ligand SMILES?

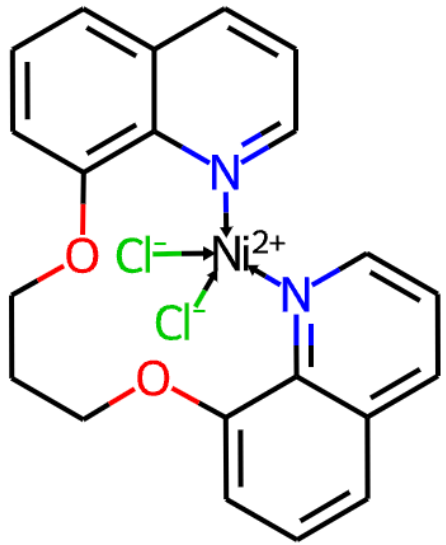


Same TM-ligand connectivity?

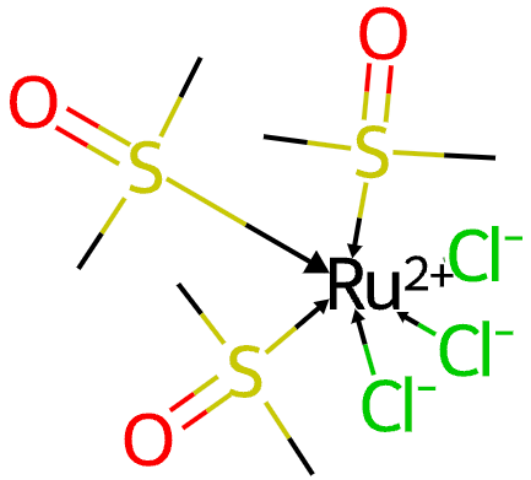




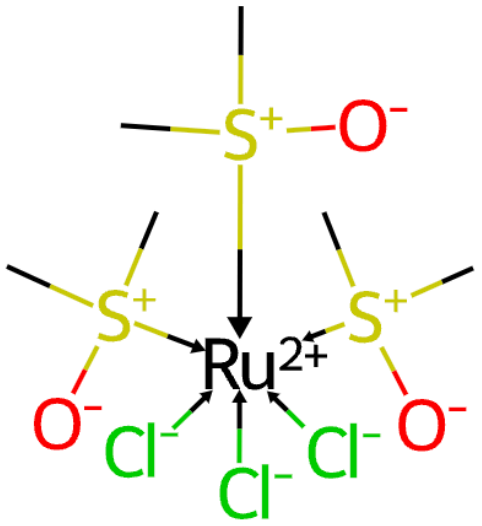
MUSQUM new CSD



MUSQUM xyz2mol



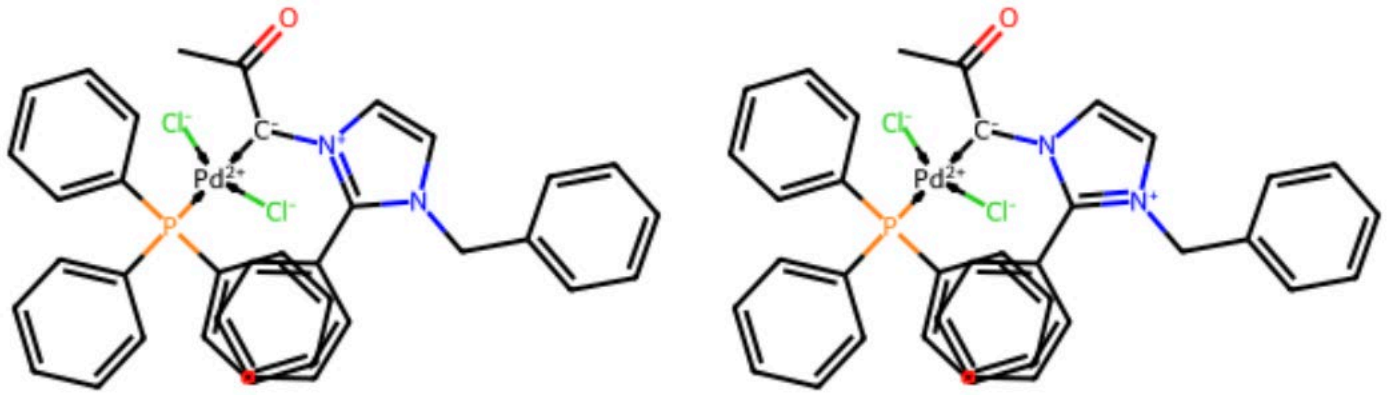
XALSEI new CSD



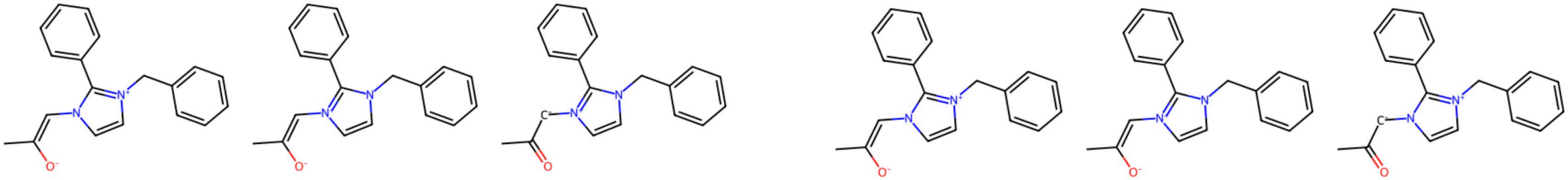
XALSEI xyz2mol

Magnus Strandgaard

We're still learning how to use ResonanceMolSupplier (RMS)

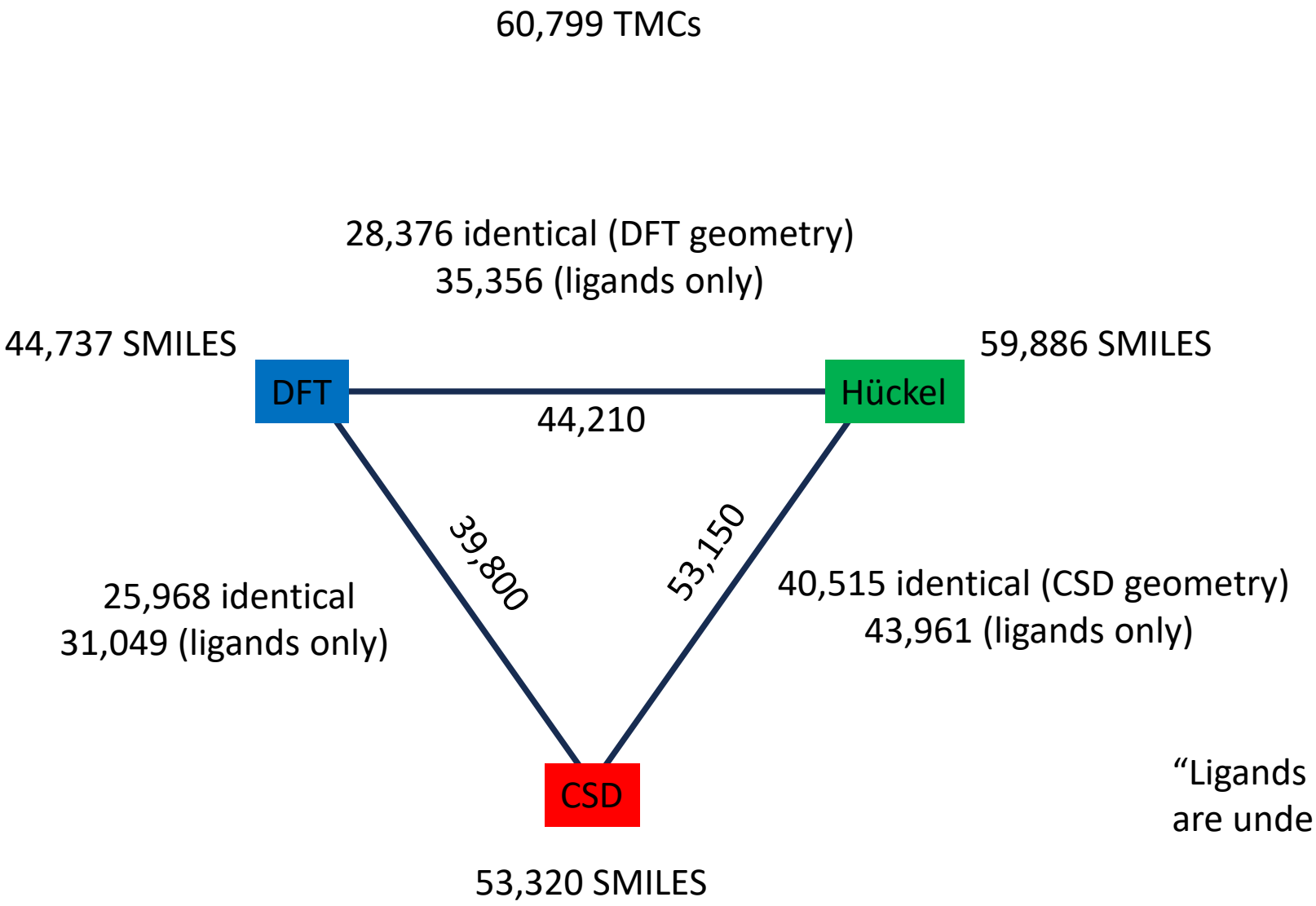


RMS on TMC



RMS on ligand

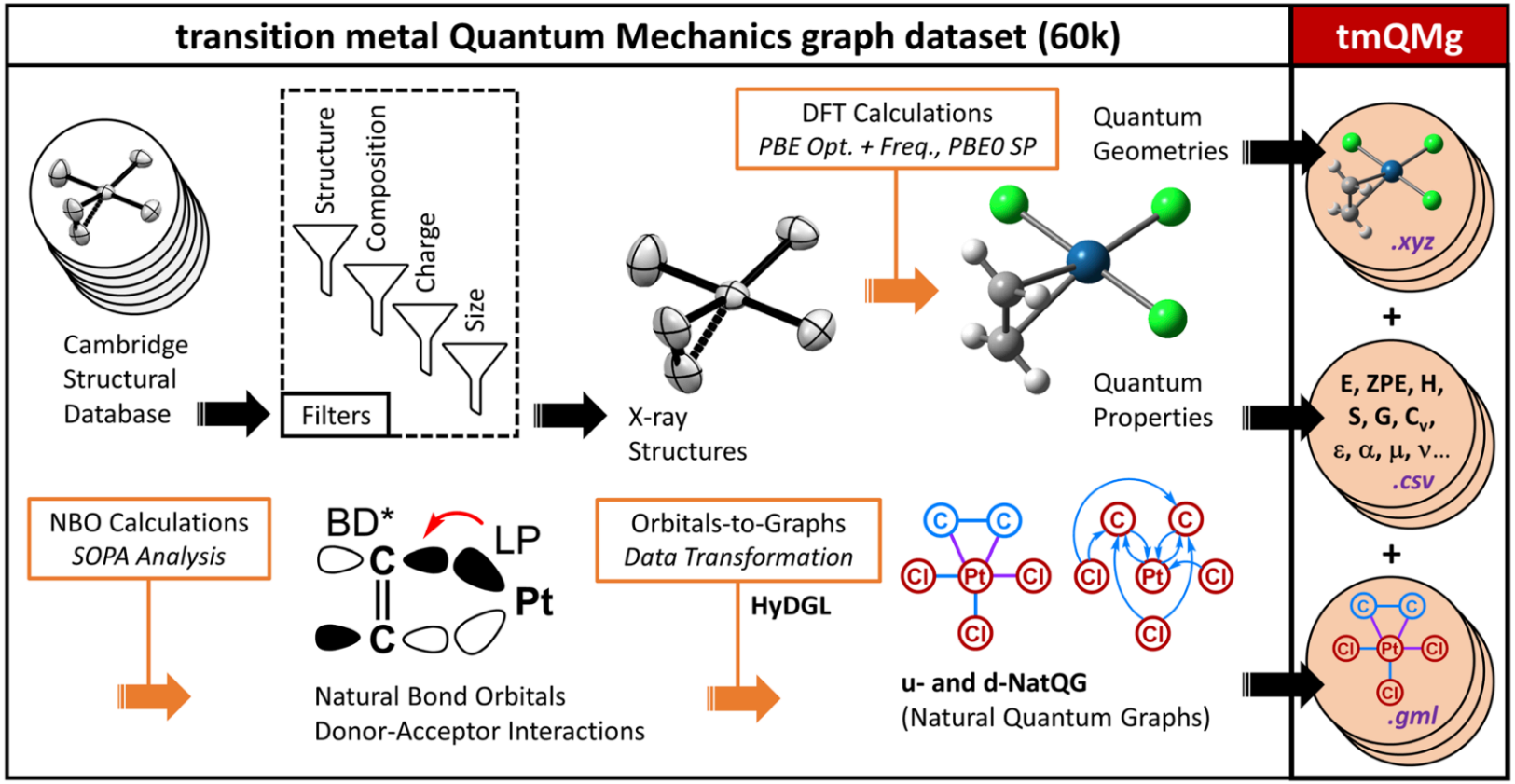
SMILES validation by comparing different approaches for Balcells 61k set



"Ligands only" numbers are underestimated

Deep learning metal complex properties with natural quantum graphs†

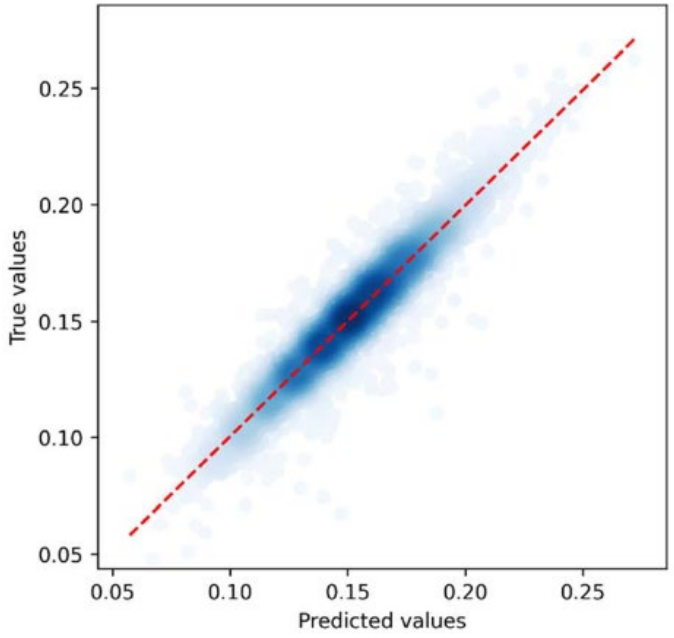
Hannes Kneiding, [a] Ruslan Lukin,[a] Lucas Lang, [a] Simen Reine, [a]
Thomas Bondo Pedersen, [a] Riccardo De Bin [b] and David Balcells *[a]



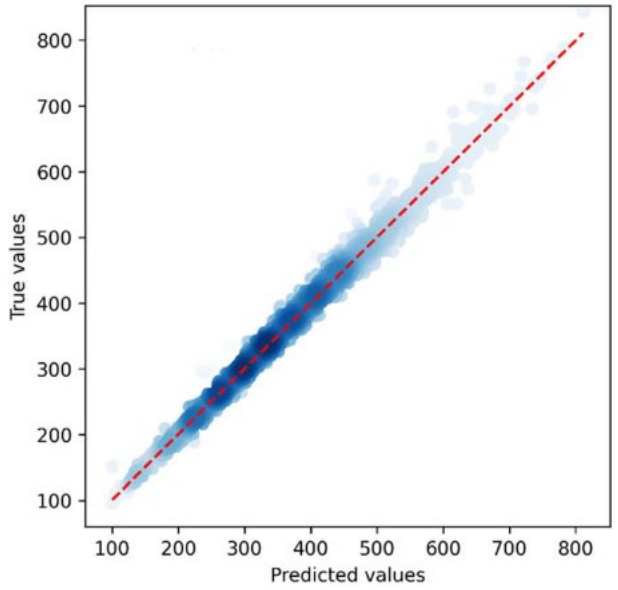
Deep learning metal complex properties with natural quantum graphs†

Hannes Kneiding, [a] Ruslan Lukin,[a] Lucas Lang, [a] Simen Reine, [a]
Thomas Bondo Pedersen, [a] Riccardo De Bin [b] and David Balcells *[a]

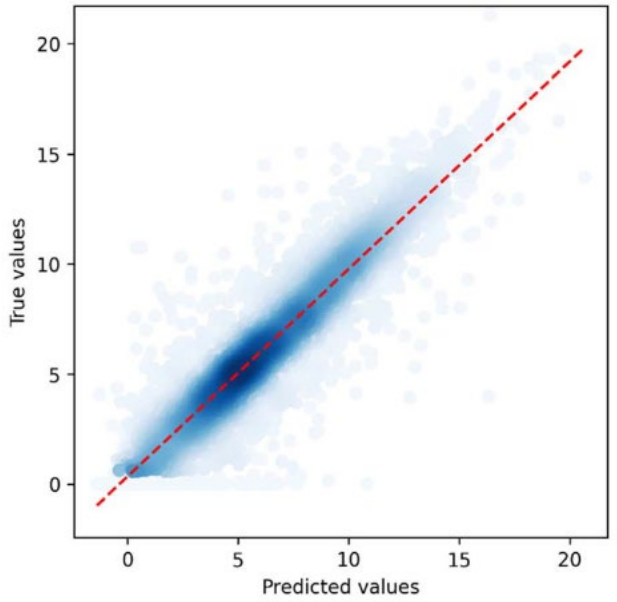
HOMO-LUMO gap



Polarizability



Dipole moment



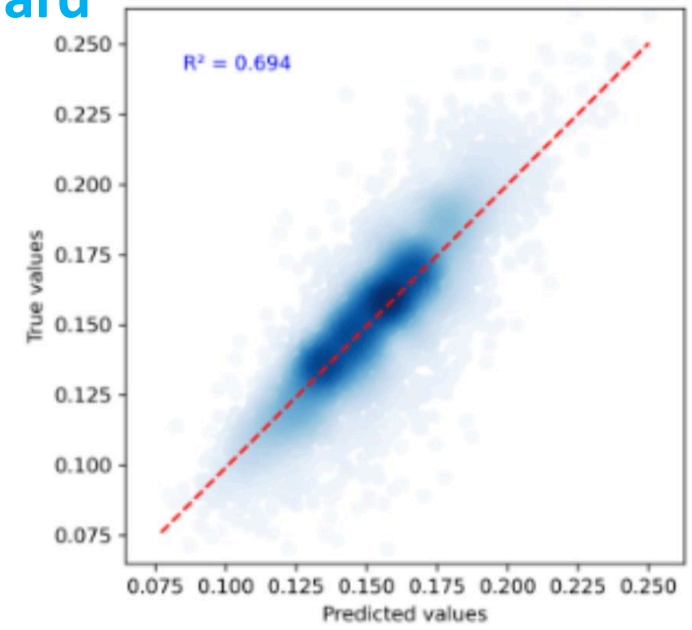
Magnus Strandgaard

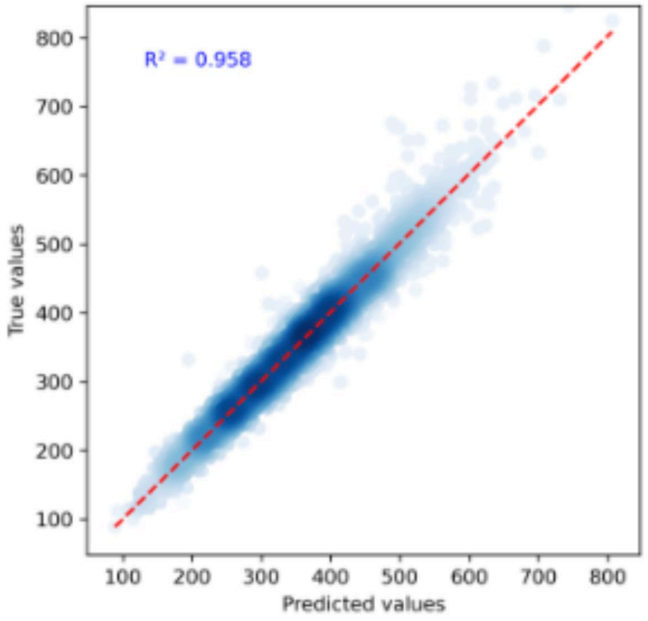
HOMO-LUMO gap

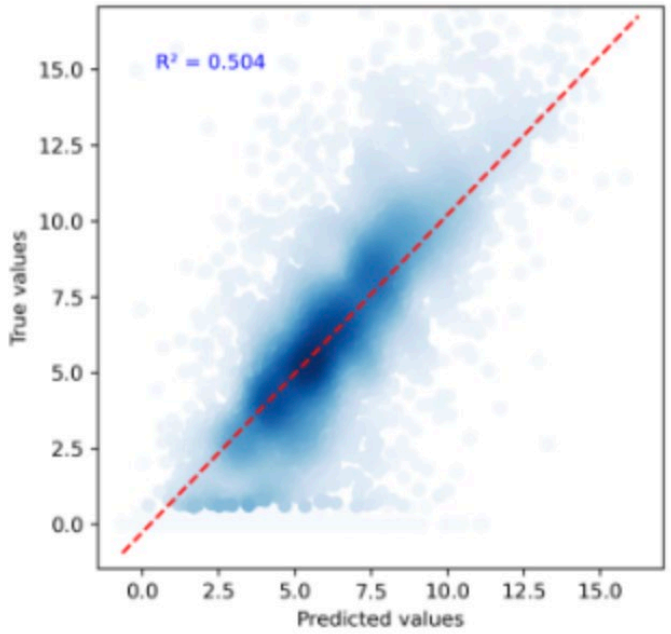
Polarizability

Dipole moment

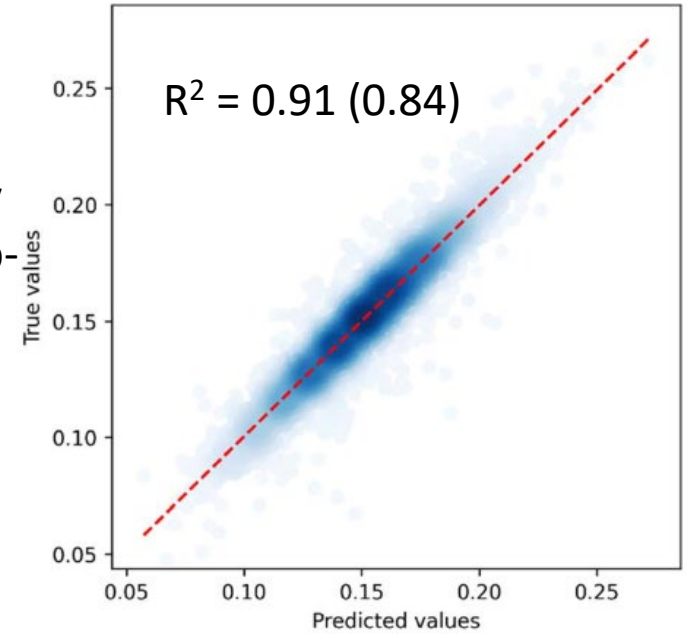
lightGBM ECFP4

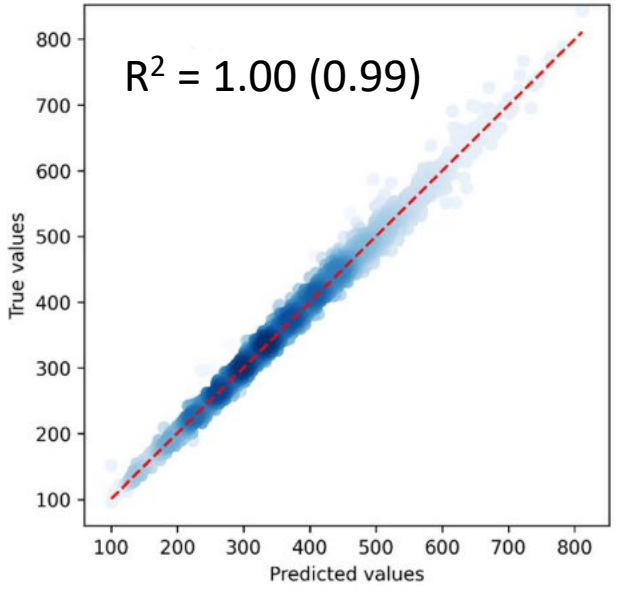


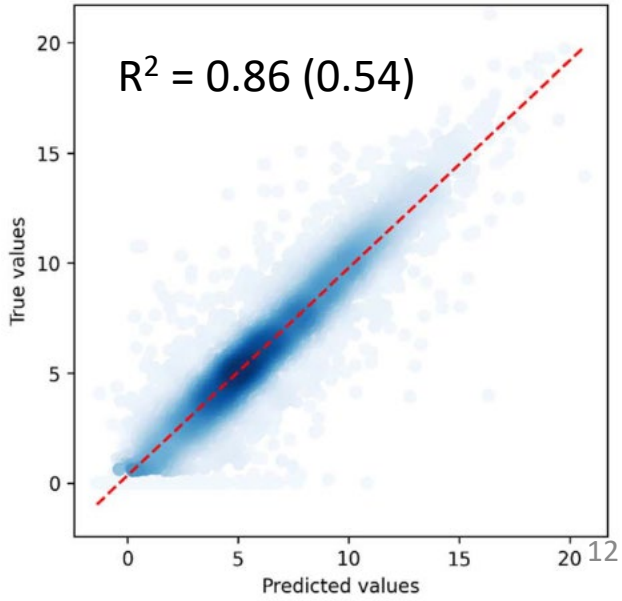




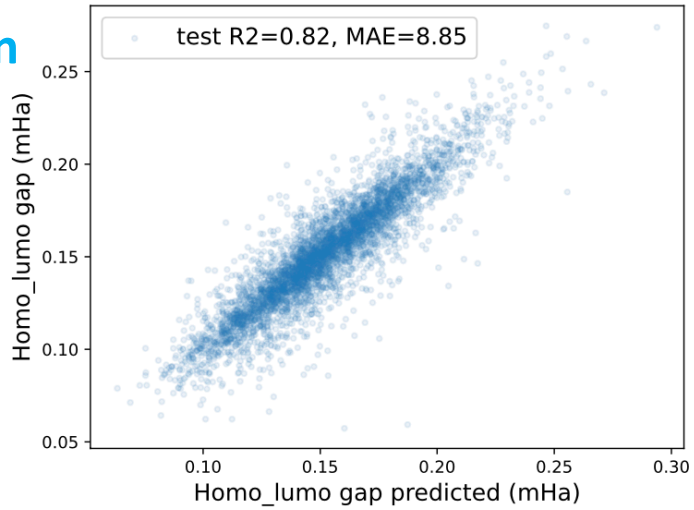
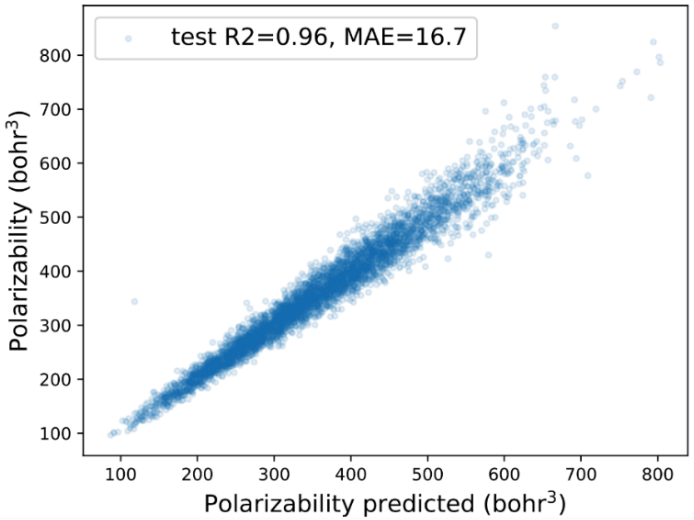
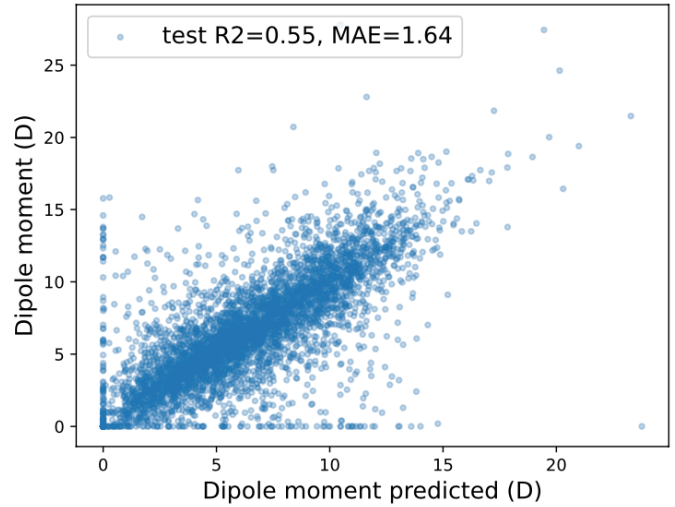
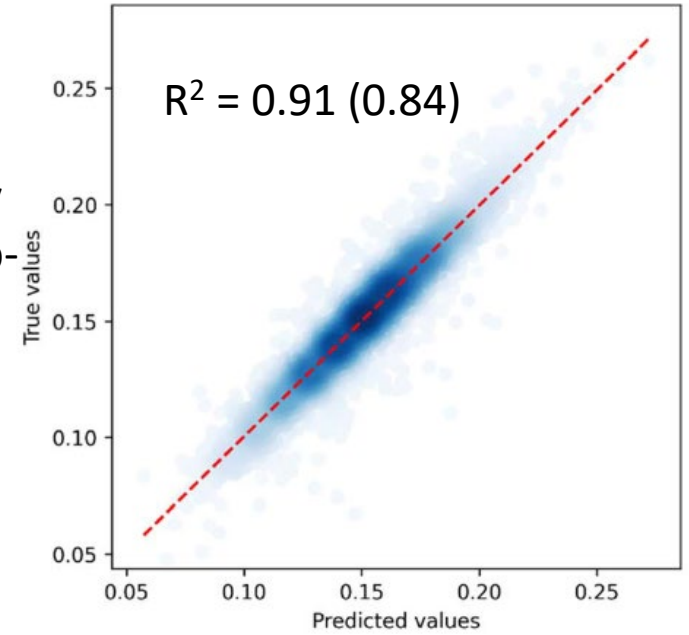
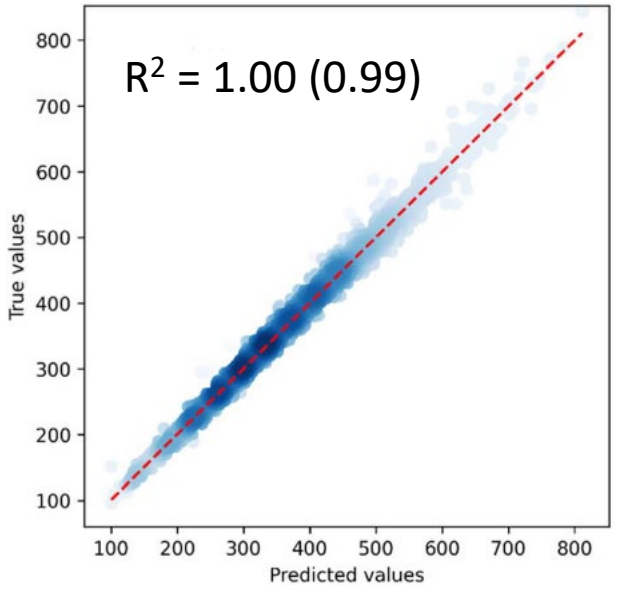
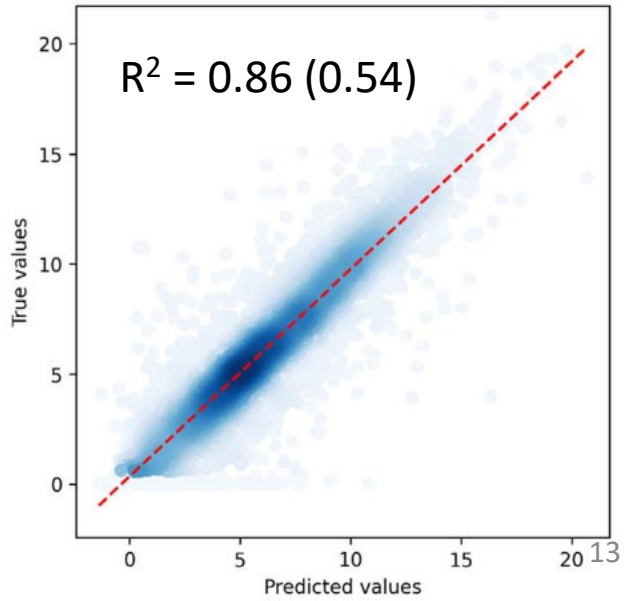
Balcells' graphConv DFT descrip-tors







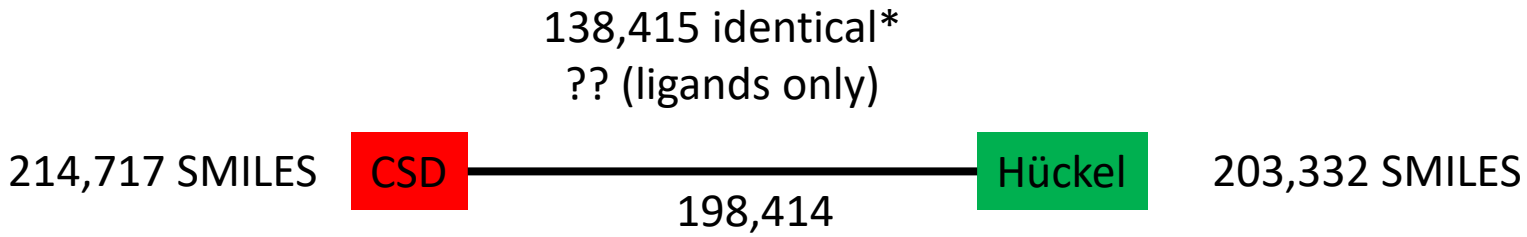
Maria Rasmussen & Laura Hemmingsen

	HOMO-LUMO gap	Polarizability	Dipole moment
RDKit-based graphConv			
Balcells' graphConv DFT descrip-tors			

Angelo Frei (York)

tmRDKit

224,656 CSD entries with a single TM



Preliminary numbers

Summary and Outlook

200k TMC dataset

xyz2mol for TMC

Hacky CCDC SMILES fixer code

Everything to be released soon + preprint

Feel free to pitch in

Hooray for

ResonanceMolSupplier 🎉 🥳

(But it still freezes up for large systems)