Supporting Information

Accurate and Rapid Prediction of pK_a of Transition Metal Complexes: Semiempirical Quantum Chemistry with a Data-Augmented Approach

Vivek Sinha*,[a], Jochem J. Laan[a] and Evgeny A. Pidko*,[a]

[a] Inorganic Systems Engineering, Department of Chemical Engineering, Faculty of Applied Sciences, Delft University of Technology, 2629 HZ, Delft, The Netherlands.

Corresponding authors:

Vivek Sinha (V.Sinha@tudelft.nl)

Evgeny A. Pidko (e.a.pidko@tudelft.nl)

Table of Contents

General remarks 3

Hessian Calculations 3

Correlation between E(A) – E(AH) and G(A) – G(AH) 4

GFN2-xTB Calculations 5

DFT//GFN2-xTB Calculations 7

DFT Calculations 9

Structure overlay plots of 75, 78 and 102 11

Structure plots of outlier complexes in Figure-SI-10.. 12

Comparison of gas phase and solvated optimized geometries 12

Ensemble of ML models learnt by AutoML 13

General remarks

Hessian Calculations

Computational cost and convergence of numerical methods are some of practical challenges in computational application of quantum chemical approaches. Identification of resource (time, CPU hours) consuming steps in the calculations and strategies to avoid them at little loss of accuracy is a promising approach to reduce the time and computer cost of quantum chemical calculations. We identify one such avenue to significantly reduce the cost of pKa calculations within GFN2-xTB and DFT calculations. Estimation of pKa requires calculation of Gibbs free energy. For both GFN2-xTB and DFT calculations Gibbs free energy can be estimated by computing the hessian matrix at the optimized geometry. Calculation of hessian is computationally intensive. An analysis of 107 DFT calculations with various sizes (N = 11 – 115 atoms) at pbe1pbe/def2-SVP level revealed that the time taken for hessian calculation grows approximately as N^3 and for N > 50 hessian calculations can consume more than 35% and for N > 80, > 50% of the total calculation time Figure-SI- 1 and Figure-SI- 2.[i]

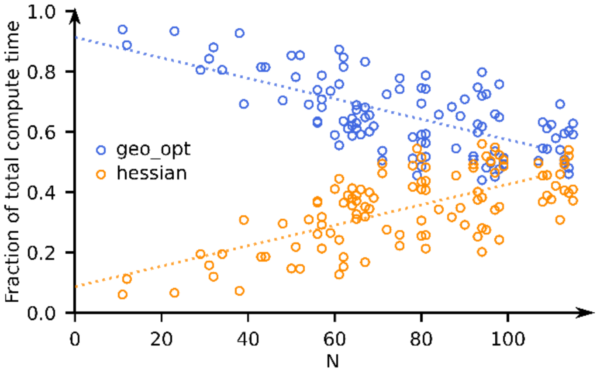


Figure-SI- 1: fraction of total computing time for geometry optimization and hessian calculations vs number of atoms (N).

Therefore, avoiding hessian calculations can help reduce the computational cost and complexity.

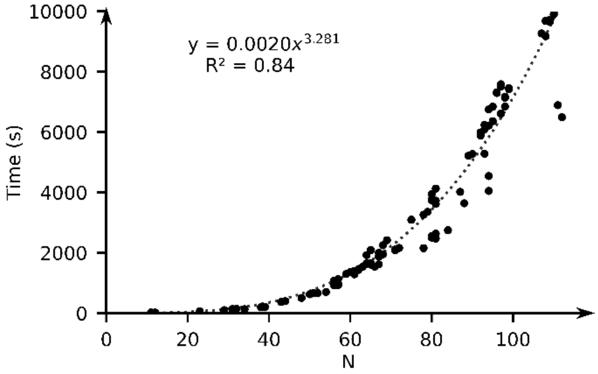


Figure-SI- 2: time taken for hessian calculation vs N.

We therefore use electronic energy to compute the PA ($\approx E(A) - E(AH)$) in our ML models.

Correlation between E(A) – E(AH) and G(A) – G(AH)

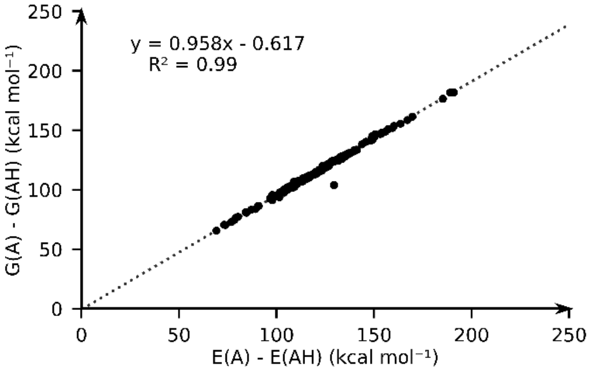


*Figure-SI- 3: Observed linear correlation between the Gibbs free energy of deprotonation versus the change in electronic free energy upon deprotonation in GFN2-xTB calculations. The outlier is complex **159** ([HFe(Py$_2$Tstacn)]$^{+2}$.*

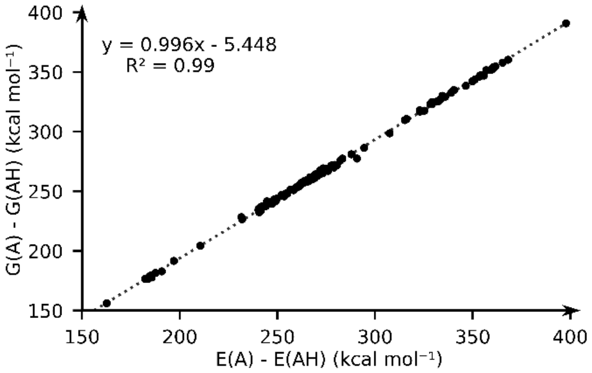


Figure-SI- 4. Observed linear correlation between the Gibbs free energy of deprotonation versus the change in electronic free energy upon deprotonation in DFT calculations.

GFN2-xTB Calculations

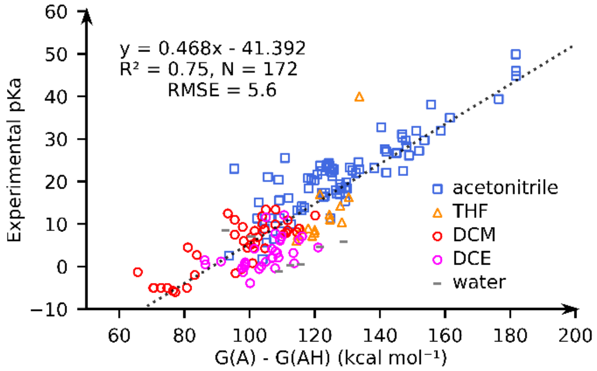


Figure-SI- 5. Experimental pK$_a$ vs PA computed as G(A) – G(AH).

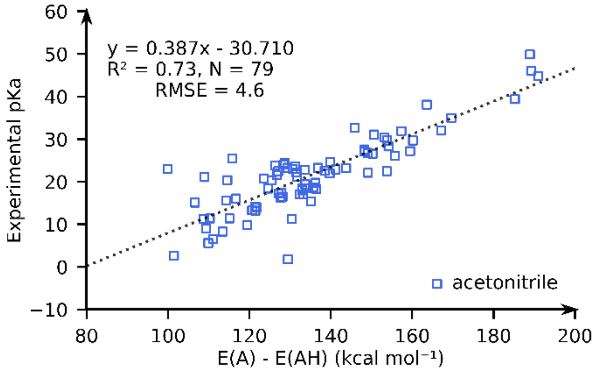


Figure-SI- 6: experimental pKa vs GFN2-xTB computed solvated PA (electronic energy) in acetonitrile.

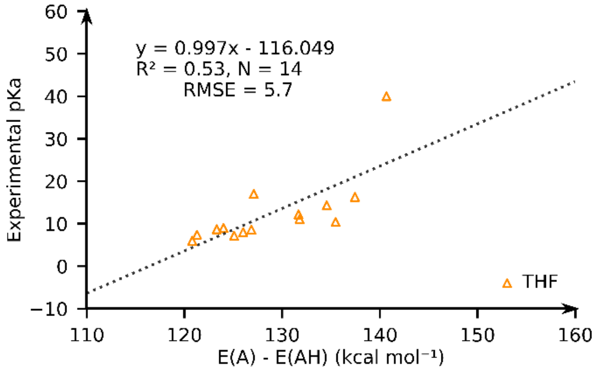


Figure-SI- 7: experimental pKa vs GFN2-xTB computed solvated PA (electronic energy) in THF.

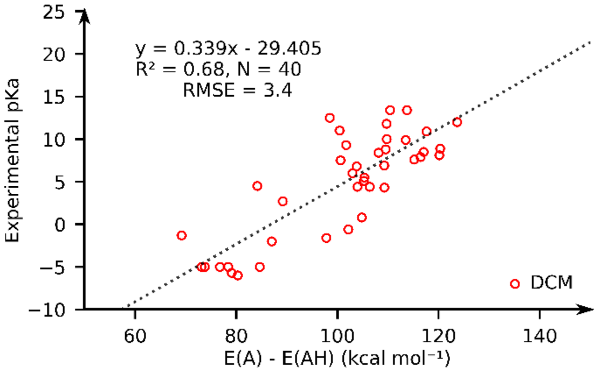


Figure-SI- 8: experimental pKa vs GFN2-xTB computed solvated PA (electronic energy) in dichloromethane.

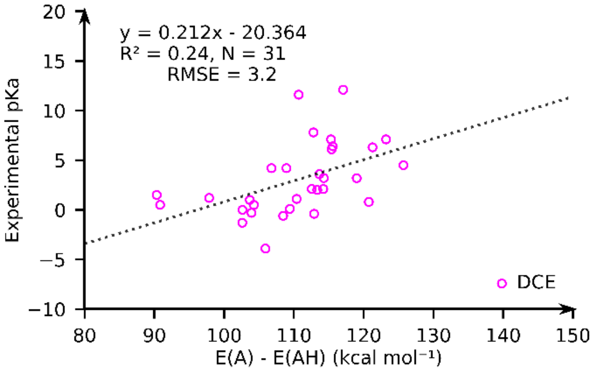


Figure-SI- 9: experimental pKa vs GFN2-xTB computed solvated PA (electronic energy) in dichloroethane.

DFT//GFN2-xTB Calculations

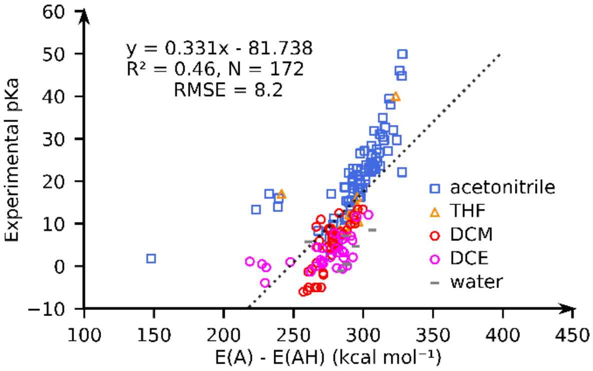


Figure-SI- 10: experimental pKa vs DFT//GFN2-xTB computed solvated PA (electronic energy).

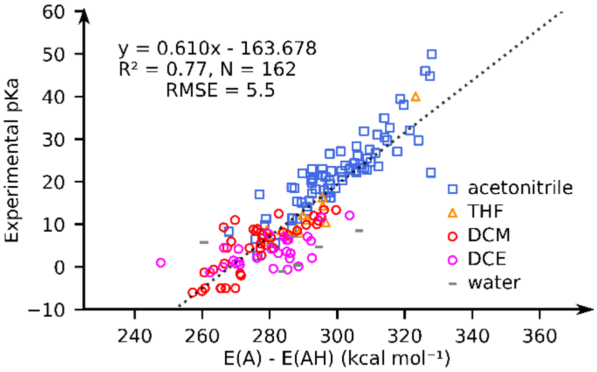


Figure-SI- 11: experimental pKa vs DFT//GFN2-xTB computed solvated PA (electronic energy), with the outliers removed.

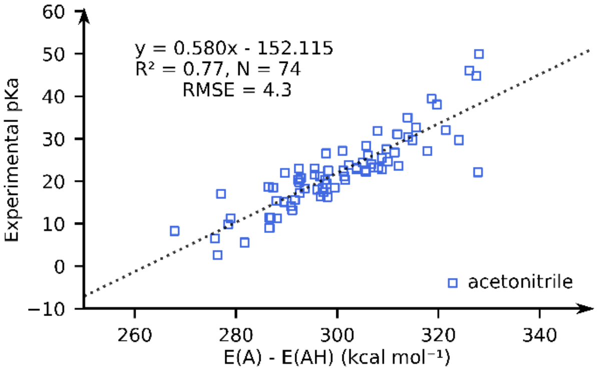


Figure-SI- 12: experimental pKa vs DFT//GFN2-xTB computed solvated PA (electronic energy) in acetonitrile.

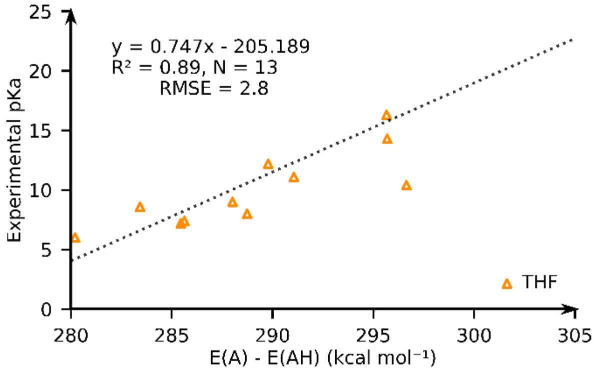


Figure-SI- 13: experimental pKa vs DFT//GFN2-xTB computed solvated PA (electronic energy) in THF.

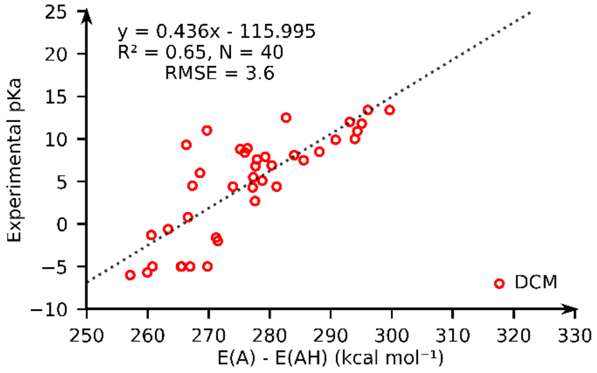


Figure-SI- 14: experimental pKa vs DFT//GFN2-xTB computed solvated PA (electronic energy) in dichloromethane.

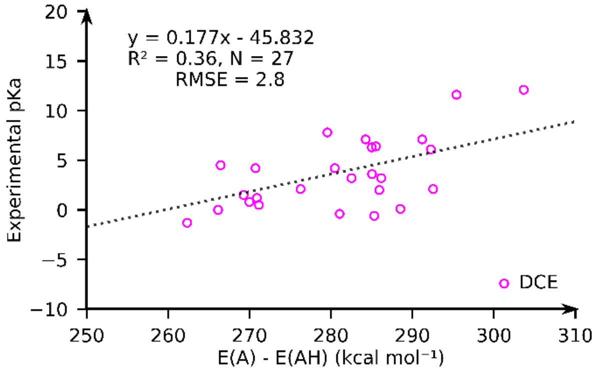


Figure-SI- 15: experimental pKa vs DFT//GFN2-xTB computed solvated PA (electronic energy) in dichloroethane.

DFT Calculations

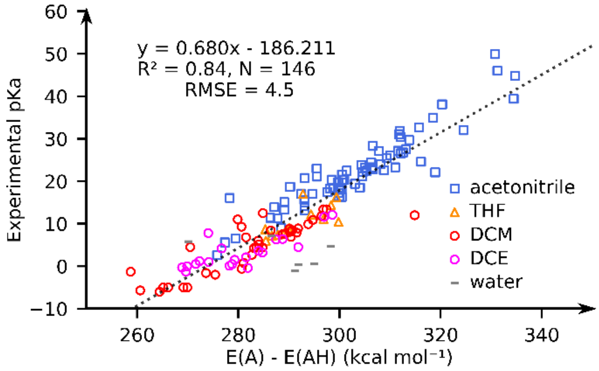


Figure-SI- 16: experimental pKa vs DFT computed solvated PA (electronic energy).

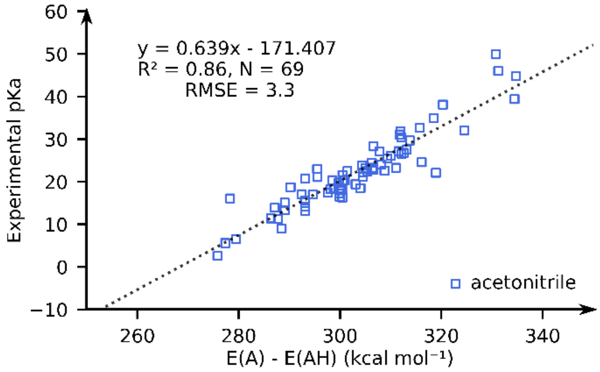


Figure-SI- 17: experimental pKa vs DFT computed solvated PA (electronic energy) in acetonitrile.

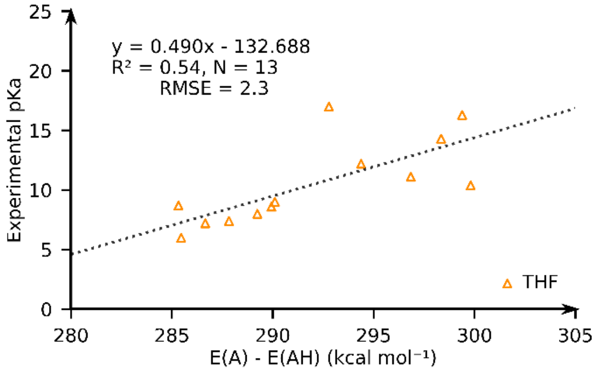


Figure-SI- 18: experimental pKa vs DFT computed solvated PA (electronic energy) in THF.

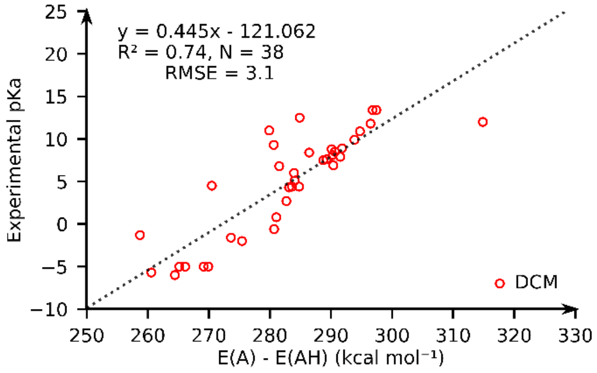


Figure-SI- 19: experimental pKa vs DFT computed solvated PA (electronic energy) in dichloromethane.

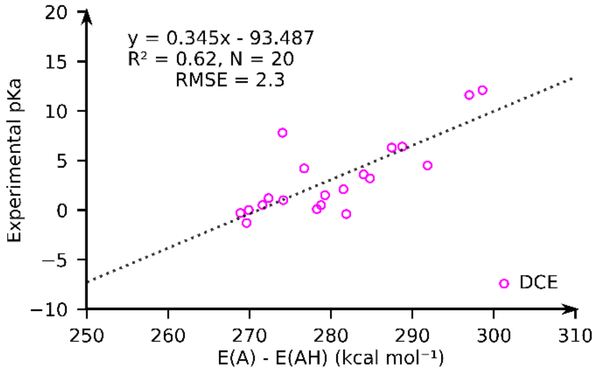


Figure-SI- 20: experimental pKa vs DFT computed solvated PA (electronic energy) in dichloroethane.

Structure overlay plots of 75, 78 and 102

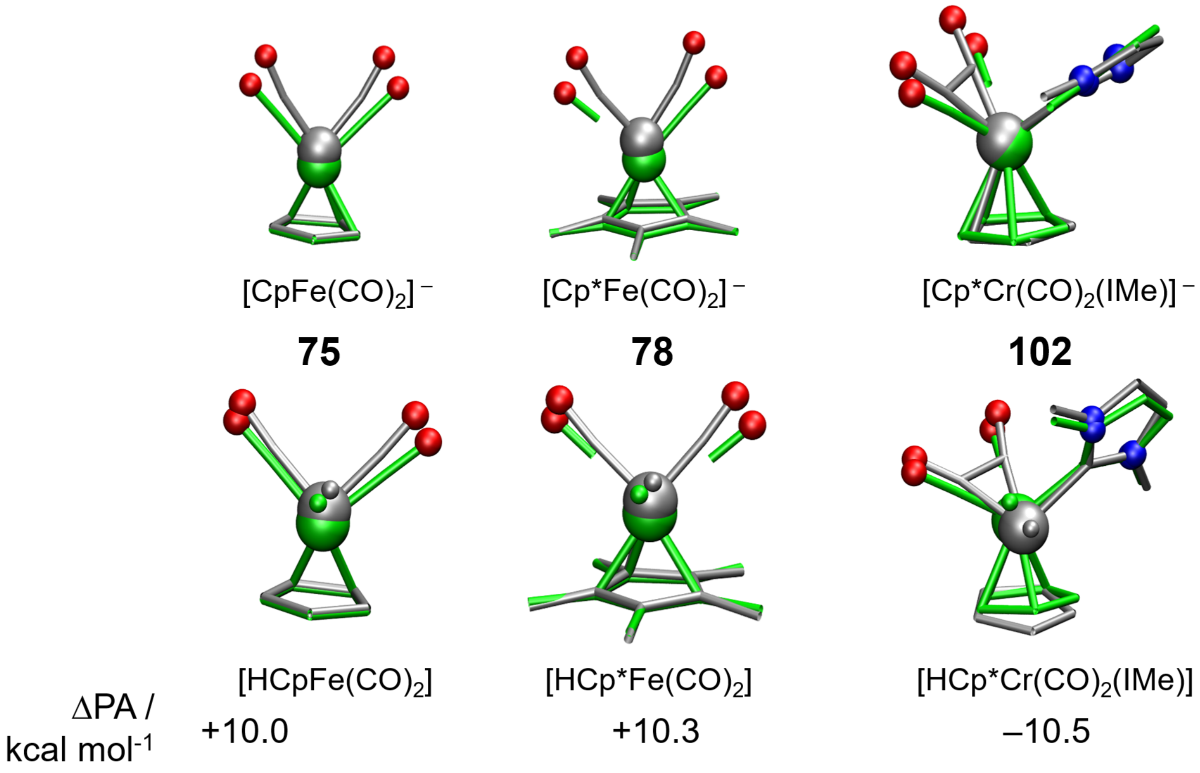
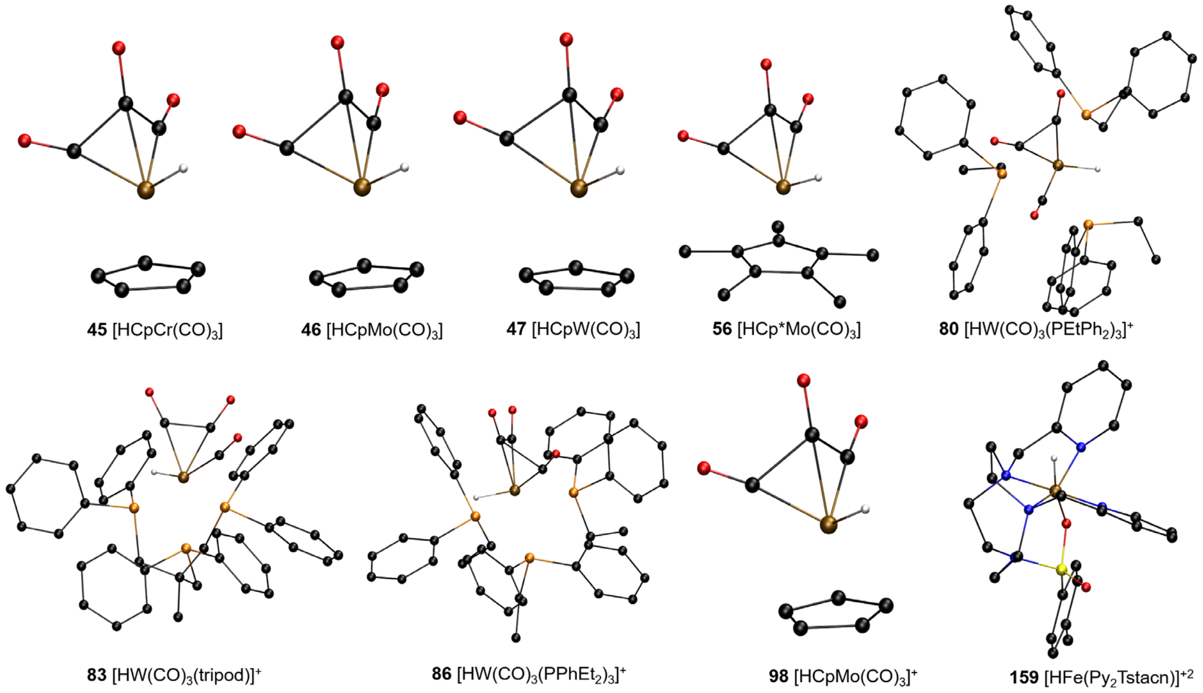


Figure-SI- 21: Structure overlay plots for conjugate base and acid forms of dicarbonyl complexes 75, 78 and 102. The C_{CO}-M-C_{CO} angle matches well with DFT predicted geometry in the acid form of complexes 75 and 78 while it is underestimated in the base form. This leads to ΔPA = 10.0 and +10.3 kcal mol^{-1}. The C_{CO}-M-C_{CO} angle is underestimated both in the acid and base forms of complex 102 leading to higher than average energy differences of the acid and base forms. The acid form also has a distorted Cp ring which attributes a ΔPA = -10.5 kcal mol^{-1}

Structure plots of outlier complexes in Figure-SI-10

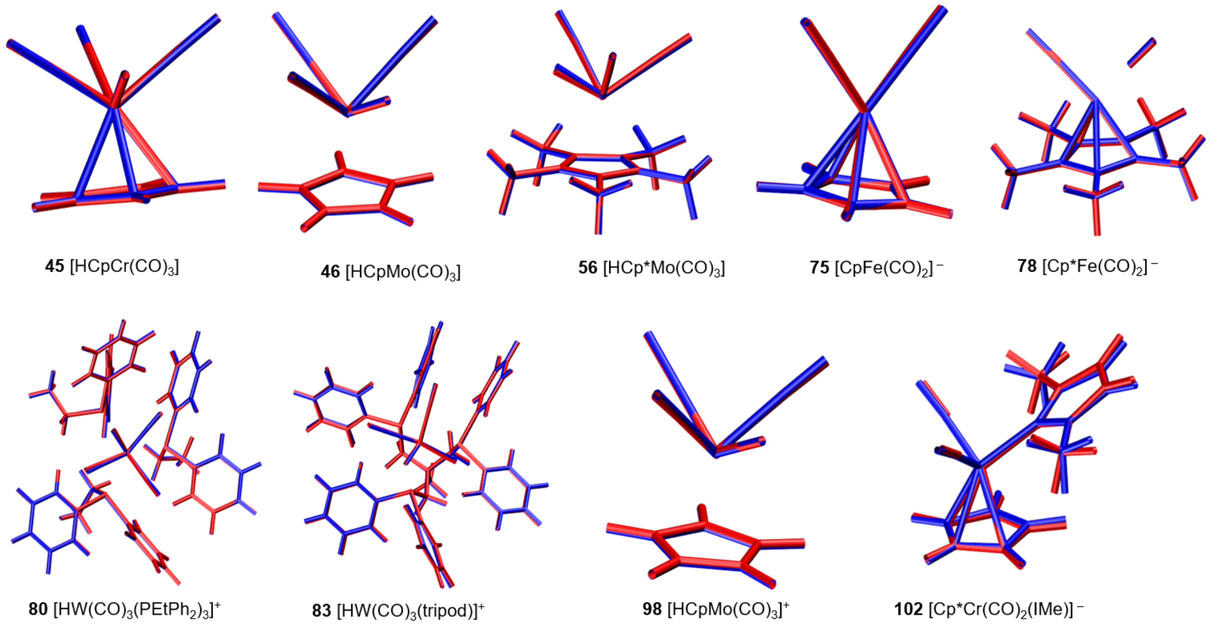


45 [HCpCr(CO)$_3$] **46** [HCpMo(CO)$_3$] **47** [HCpW(CO)$_3$] **56** [HCp*Mo(CO)$_3$] **80** [HW(CO)$_3$(PEtPh$_2$)$_3$]$^+$

83 [HW(CO)$_3$(tripod)]$^+$ **86** [HW(CO)$_3$(PPhEt$_2$)$_3$]$^+$ **98** [HCpMo(CO)$_3$]$^+$ **159** [HFe(Py$_2$Tstacn)]$^{+2}$

Figure-SI- 22. Structure plots of acid form of complexes that are outliers in the plot between exp. pK$_a$ and DFT//GFN2-xTB computed PA in Figure-SI- 10.

Comparison of gas phase and solvated optimized geometries



45 [HCpCr(CO)$_3$] **46** [HCpMo(CO)$_3$] **56** [HCp*Mo(CO)$_3$] **75** [CpFe(CO)$_2$]$^-$ **78** [Cp*Fe(CO)$_2$]$^-$

80 [HW(CO)$_3$(PEtPh$_2$)$_3$]$^+$ **83** [HW(CO)$_3$(tripod)]$^+$ **98** [HCpMo(CO)$_3$]$^+$ **102** [Cp*Cr(CO)$_2$(IMe)]$^-$

Figure-SI- 23. DFT optimized structure plots of selected complexes. Geometries optimized in the presence of solvation potential (SMD) are shown in blue while the geometries optimized in the gas phase are shown in red. The gas phase and solvated geometries overlap well showing that the impact of solvation potential on the optimized geometries is minimal.

Ensemble of ML models learnt by AutoML

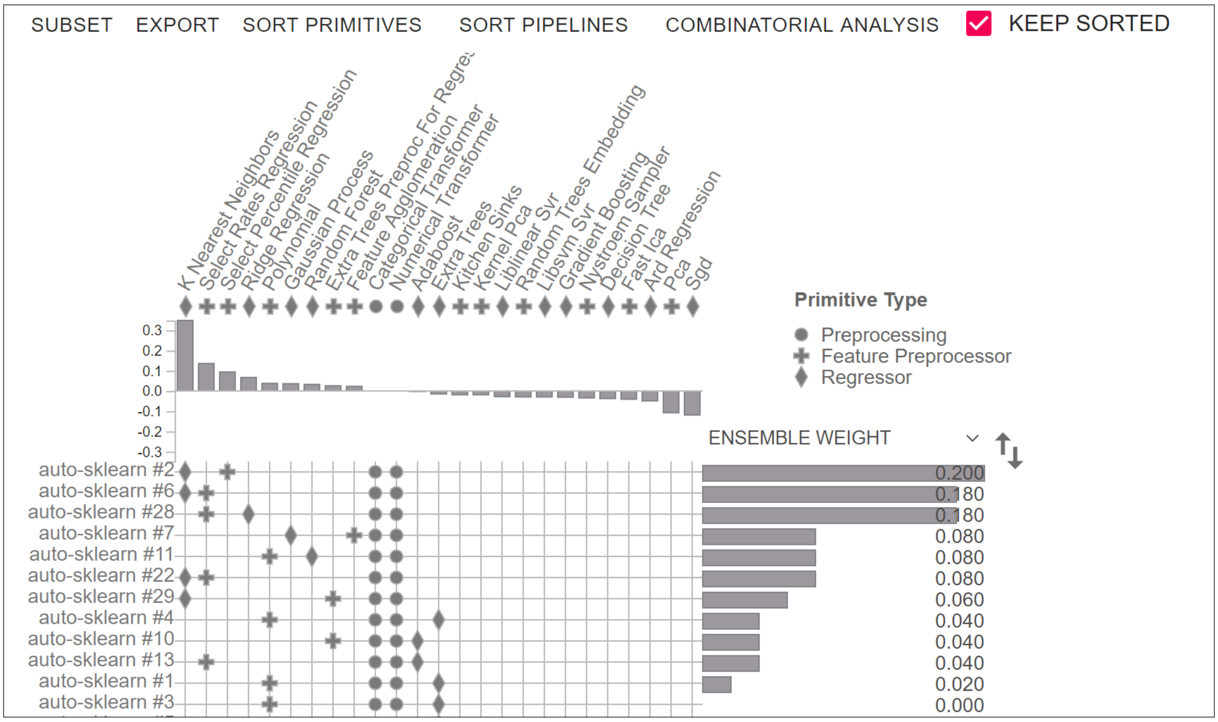


Figure-SI- 24. Snapshot of the output of PipelineProfiler[ii] on the ensemble of ML models learnt by the Auto-sklearn method.

[i] This also depends on several factors such as the optimizer used, initial geometry etc.

[ii] arXiv:2005.00160v2