Supplementary Information (SI) for Digital Discovery.

Supporting Information for: Towards a Universal Scaling Method for Predicting Equilibrium Constants of Polyoxometalates

Jordi Buils[a,b], Diego Garay-Ruiz[a], Enric Petrus[a,c], Mireia Segado-Centellas[b], and Carles Bo[a,b]

[a]Institute of Chemical Research of Catalonia (ICIQ), The Barcelona Institute of Science and Technology (BIST), Av. Països Catalans, 16, 43007 Tarragona, Spain.
[b]Departament de Química Física i Inorgànica, Universitat Rovira i Virgili, Marcel·li Domingo s/n, 43007, Tarragona, Spain.
[c]EAWAG: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133 Dübendorf 8600, Switzerland

January 24, 2025

Contents

1 Validation of experimental formation constants **3**

2 Universality across functionals **5**

3 Speciation of AsMo with original intercept **6**

4 Feature sets for multi-linear regression **7**

5 Validation of intercept predictions **8**

6 Phase diagrams for arsenomolybdates **10**

1 Validation of experimental formation constants

Throughout the characterization of the universality of the scaling parameter across the previously known systems, we observed that the Mo system seemed to strongly depart from the rest, with a slope of *0.22* that deviated from the expected value of around *0.3* calculated for the rest of the systems. Consequently, including Mo-based species in the global regression did also strongly modify the slope parameter of the overall regression. Nonetheless, the foundational studies from the monomer to the octamolybdate[1] did not show this discrepancy.

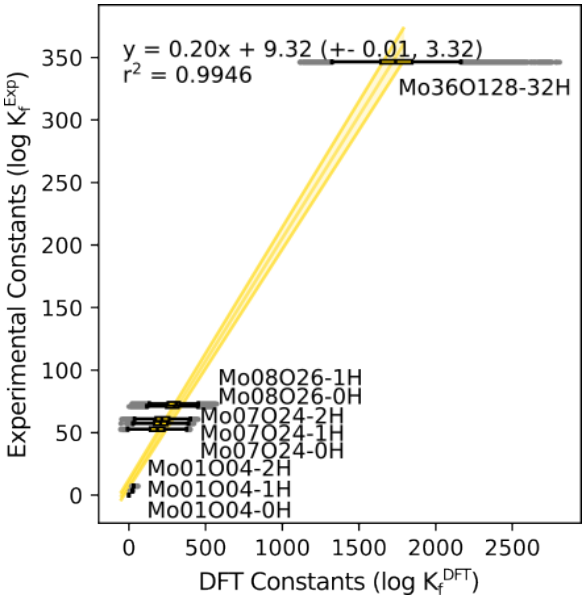

Figure 1: Box-and-whisker plot representation for formation constants in Mo system[2].

From this information and from Figure 1, the very large Mo_{36} cluster seemed to be the most likely reason for the mismatch. In order to properly explore its effect on our overall findings, we designed a leave-one-out strategy in which we carried out the linear regression on the median K_f values of all complexes in the dataset except for one, plotting the distribution of intercepts against the slope, as well as the correlation coefficient against the slope (Figure 2).

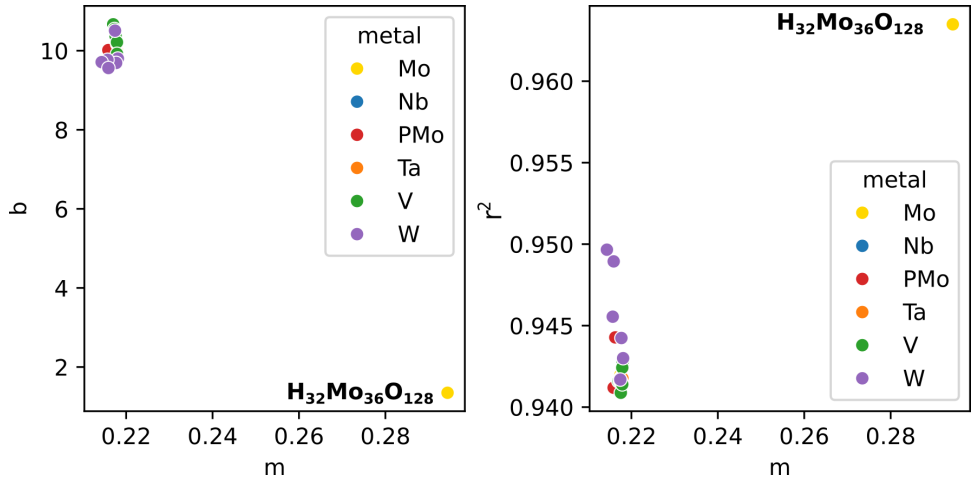

Figure 2: Regression parameter mapping from leave-one-out strategy, plotting intercept against slope (left) and r^2 against slope (right)

.

We can observe how the spread in all parameters for most species is minimal, except for the case of Mo_{36}. Removing this species immediately switches the trend, recovering the expected

slope of $m \approx 0.3$, while all other observations remain around $m = 0.22$. Discarding Mo_{36} does also result in a remarkable improvement of the quality of the regression, with r^2 reaching 0.965. Due to the very large value of this formation constant, it sits in a region where no other species in the dataset appear, and therefore possible inaccuracies on the experimental constant K_{exp} cannot be mitigated. Consequently, we decided to drop Mo_{36} from the set of experimental constants used for fitting: all results collected in the main text have been collected under this assumption. A somehow similar procedure was proposed by Uzal-Varela et al.[3], to correct the stability constants of reported gadolinium complexes.

2 Universality across functionals

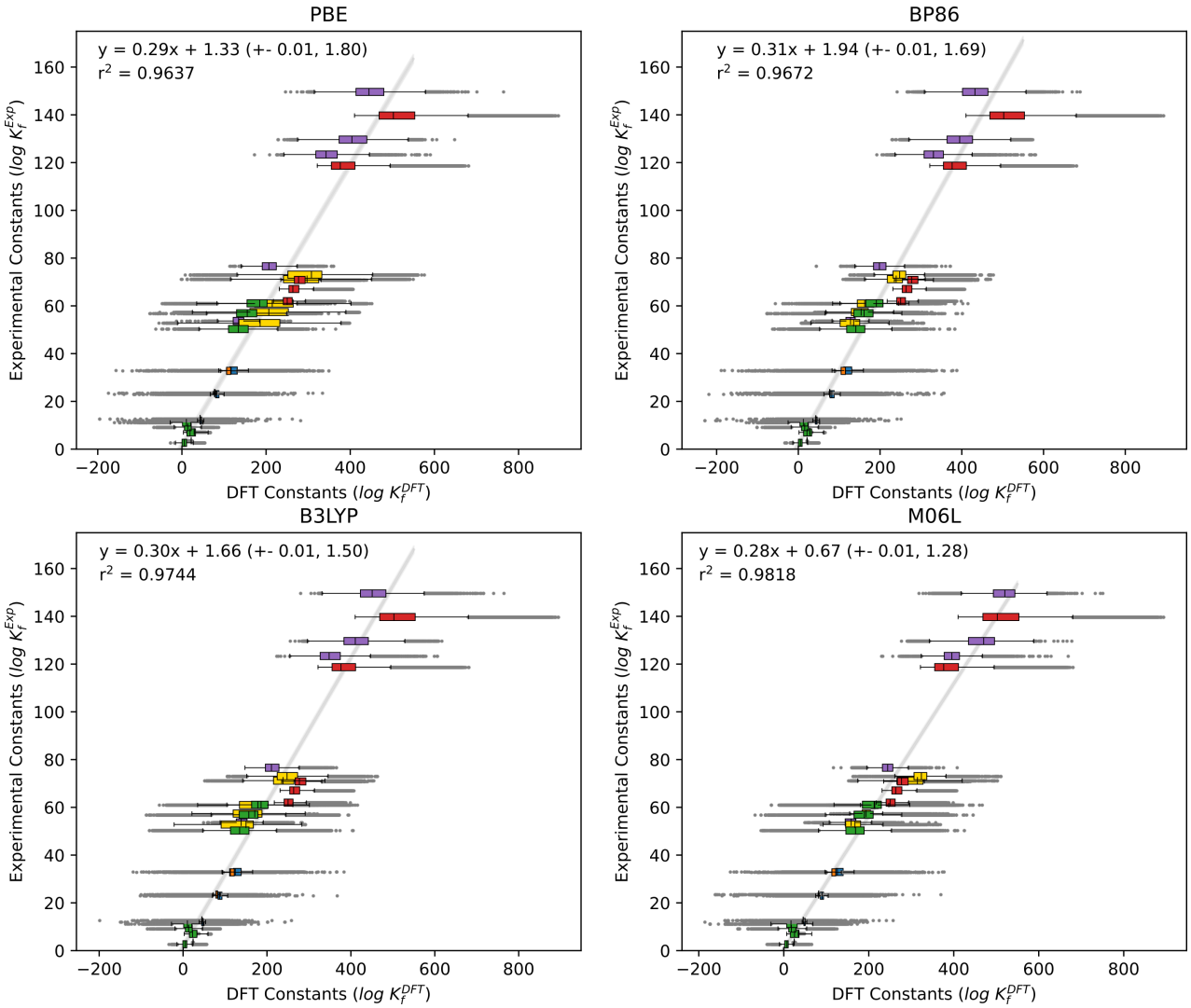

Figure 3: Common box-and-whisker plot representation for formation constants in Mo, W, V, Nb, Ta and PMo systems, throughout the four tested functionals: PBE, BP86, B3LYP and M06L. Y-axis corresponds to experimental values reported in the literature[2,4–8].

3 Speciation of AsMo with original intercept

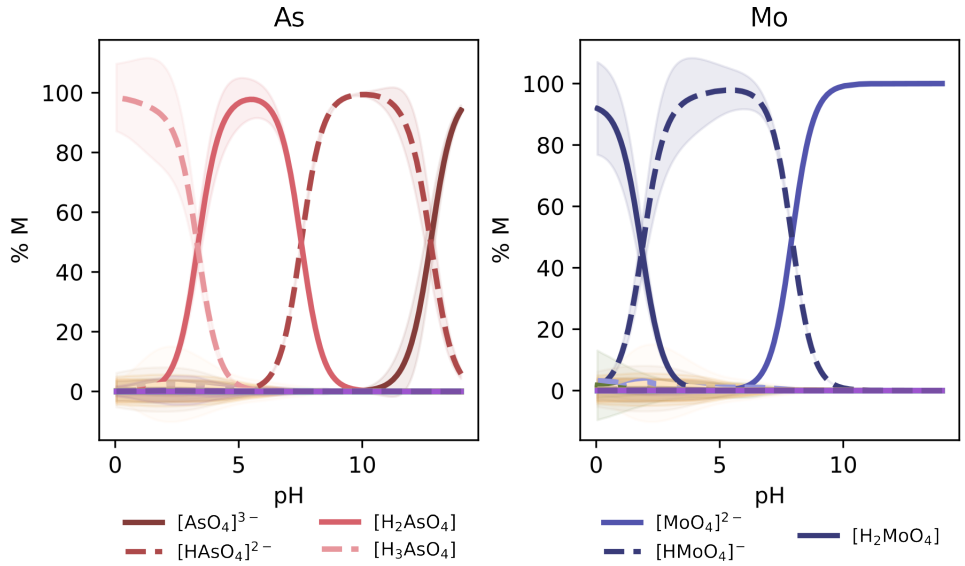

Figure 4: Speciation diagrams for arsenomolybdates considering percentage of As (left) and Mo (right), directly using the scaling parameters from Equation 2 in main text, averaging around 10^4 speciation models.

Figure 4 showcases the problems associated with employing the intercept in Equation 2 in the main text ($K_{exp} = 0.29 \cdot K_{DFT} + 1.33$): with this intercept, only monomers are observed, without any trace of the larger arsenomolybdate clusters expected from the experimental data[9].

4 Feature sets for multi-linear regression

As discussed in the main text, the six previously studied POM systems (Mo, W, V, Nb, Ta and PMo) were used as the training set to predict the intercept for the AsMo system, known to be $b_{AsMo} = -7.99$.

Features	N	R^2	b_{pred}	$\lvert b_{err}\rvert$
(Q_1, Q_3, max, range)	4	0.978	-7.81	0.18
(min, Q_3, max, range)	4	0.984	-7.63	0.36
(mean, Q_3, max, range)	4	0.998	-8.44	0.45
(min, Q_1, Q_3, max)	4	0.898	-8.63	0.64
(mean, Q_1, max, range)	4	0.948	-8.68	0.69
(Q_3, max, range)	3	0.967	-7.29	0.7
(std, Q_3, max, range)	4	0.967	-7.29	0.7
(mean, max, range)	3	0.948	-8.72	0.73
(mean, max, range, IQR)	4	0.948	-8.73	0.74
(mean, min, max, range)	4	0.955	-8.74	0.7

Table 1: Selection of the 10 best feature combinations regarding the error (in absolute value) in the prediction of the test set (AsMo), with target value b = -7.99. Entries in bold correspond to sets with a smaller number of features. Values in the Table sorted in decreasing order respect to the $|b_{err}|$ value.

Table 1 collects the ten combinations whose prediction of the test intercept is closer to the expected value: from these, we proceeded to select the two that only comprised three distinct features to reduce the risk of overfitting.

5 Validation of intercept predictions

Despite the moderate values of the error between predicted and expected intercepts for each system in the cross-validation procedure (as collected in the heatmap in Figure 6 of the main text), we were also interested in analyzing the limit situations: the combinations of systems where the predictions were farther from the expected value.

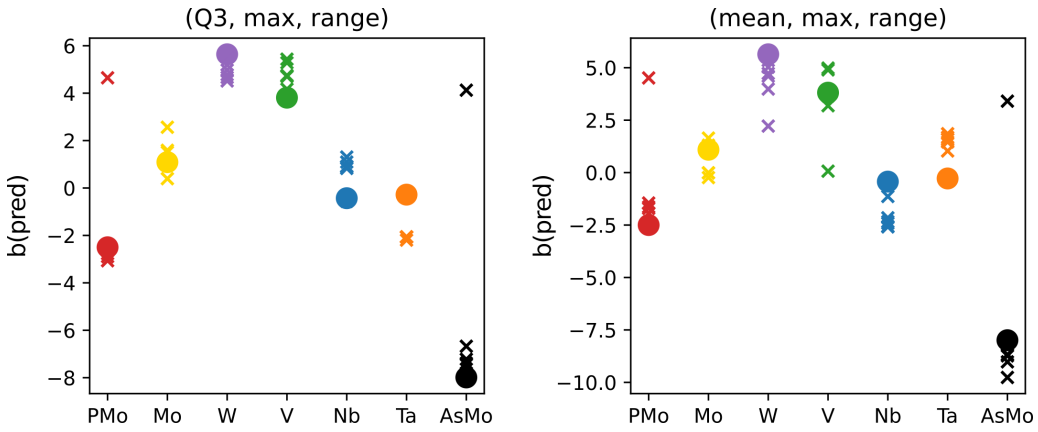

Figure 5: Dispersion of predicted intercepts b through the leave-two-out cross-validation procedure for the two three-feature combinations selected for further analysis. Circles correspond to the target intercept values, while crosses represent the predicted values.

Figure 5 shows all b predictions in the dataset: for a given POM system, there are six entries corresponding to all six combinations where that system is included in the test set together with each of the other systems. This illustrates how for both descriptor sets, the dispersion of predicted intercepts for IPAs is moderate, while the two HPAs have two clear outliers corresponding to the (PMo, AsMo) combination: that is, the situation where no information about heteropolyanionic systems was included in the training set. In direct consequence, this misprediction, far from being worrying, actually demonstrates how the proposed multi-linear regression strategy is capturing the chemical nature of the treated systems, despite its simplicity.

Moreover, we carried out a deeper exploration for the *(Q_3, max, range)* descriptor set used to develop the final predictive model. First, we compared the relative errors between model predictions and the target intercept values in three situations: i) model trained with the seven systems, ii) leave-one-out strategy, predicting the species for which the model is not trained, and iii) leave-two-out strategy, predicting all pairs of species out of training. For this case, we selected the largest error from the set of six predictions for each species, except for AsMo and PMo, where the outliers shown in Figure 5 when the training set lacks any heteropolyanionic system were removed.

The errors showcased in Figure 6 are small and quite consistent, even in the strictest case where only five systems are used for training and the worst prediction for each system is selected. Yet again, this confirms the solidity of our approach to predict the expected values of the intercepts without any experimental data. It is also remarkable that both Nb and Ta show larger errors in the baseline 0-out model, which might be traced back to the reduced number of experimental formation constants used for the characterization of the target intercepts[10] compared with the rest of the systems. Furthermore, we also explored the distribution of the multi-linear regression coefficients through a more strict leave-three-out validation procedure, further reducing the size of the training set (Figure 7).

These results confirm the robustness of the approach: the coefficients c_1 to c_3 show minimal variations, with the only major outlier for all three of them being the (Ta, AsMo, PMo) model.

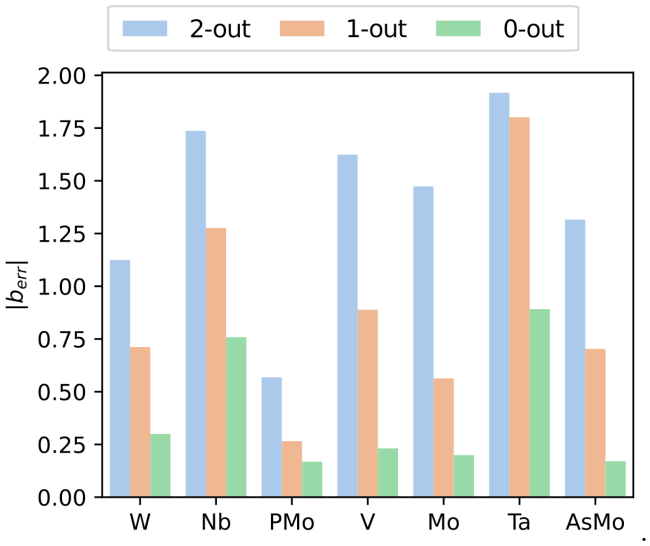

Figure 6: Relative errors of 0-out (green), 1-out (orange) and 2-out (blue) cross-validation procedures for the *(Q_3, max, range)* descriptor set .

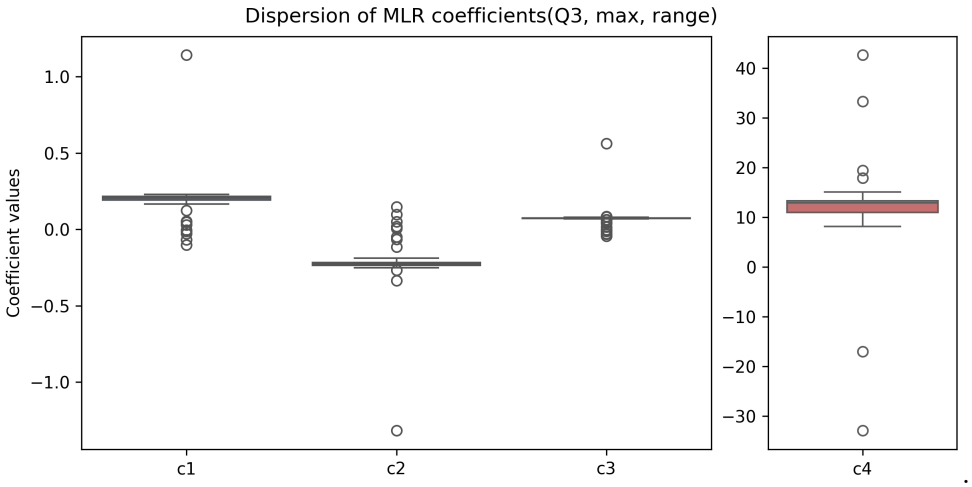

Figure 7: Box-and-whiskers representation of the distribution of c_1, c_2, c_3 and c_4 coefficients for the multilinear regression using the (Q_3, max, range) descriptor set.

For the rest, even most of the situations where only four systems are retained in the training set remain quite consistent, again highlighting the adequacy of the strategy. While the independent term c_4 shows a much larger degree of variation, the major outliers are combinations of three dropped features, including at least one of the HPAs (if not both) as well as V, which due to its amphoteric character is also expected to be remarkably different from the rest.

Consequently, despite the very small sample size (seven systems), these findings confirm that the proposed predictive model is revealing important chemical features of the systems under study, and that it can be used as a part of an universal scaling for the formation constants of polyoxometalates.

6 Phase diagrams for arsenomolybdates

Apart from the speciation diagrams (Figure 7 of the main text) and in line with our study on phosphomolybdates[11], we used the same set of SMs selected for the speciation with the 1:1 ratio to compute the corresponding phase diagrams, plotting the major species at varying $[As]/[Mo]$ ratios. It shall be noted that we consider the main species for arsenic and the main species for molybdenum, resulting in two plots (Figure 8).

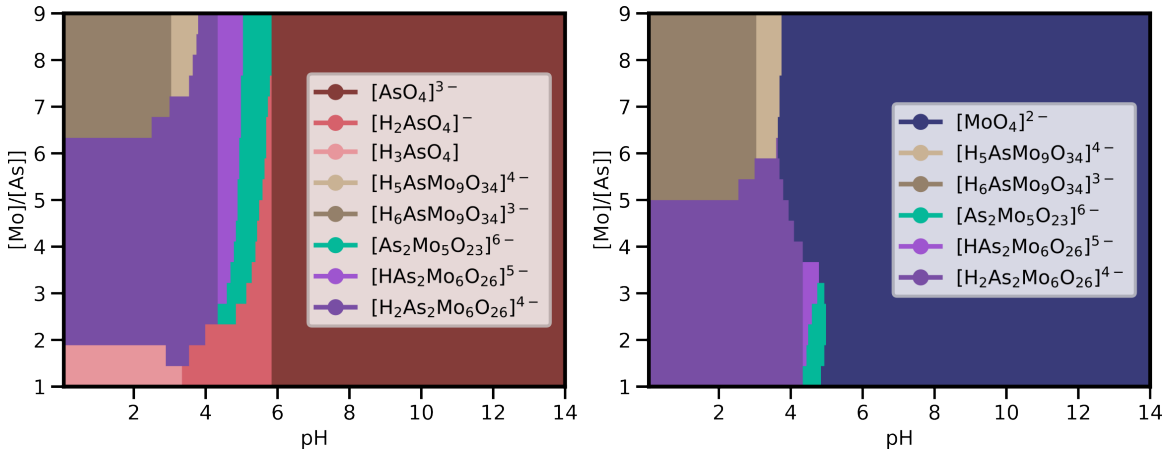

Figure 8: Phase diagrams for AsMo system, highlighting the major species for As (left panel) and Mo (right panel). A total of 18 As:Mo ratios were used, determining the average speciation of the $\approx 6 \cdot 10^4$ speciation models selected throughout the statistical pipeline.

These phase diagrams showcase the same trends observed for the speciation, with $\{As_2Mo_6\}$ dominating at low-to-moderate heteroatom/metal ratios and acidic pH, and $\{AsMo_9\}$ taking over as the proportion of molybdenum increases. From $pH = 6$ onwards, nucleation is suppressed and only monomers dominate the speciation. Moreover, the $\{As_2Mo_5\}$ becomes the dominant species for As in a narrow region around $pH = 5$: there, at large ratios, most of the molybdenum remains as $[MoO_4]^{2-}$.

References

[1] E. Petrus, M. Segado and C. Bo, *Chem. Sci.*, 2020, **11**, 8448–8456.

[2] J. Cruywagen, *Advances in Inorganic Chemistry*, Elsevier, 1999, vol. 49, pp. 127–182.

[3] R. Uzal-Varela, A. Rodríguez-Rodríguez, H. Wang, D. Esteban-Gómez, I. Brandariz, E. M. Gale, P. Caravan and C. Platas-Iglesias, *Coord. Chem. Rev.*, 2022, **467**, 214606.

[4] G. M. Rozantsev and O. I. Sazonova, *Russ. J. Coord. Chem.*, 2005, **31**, 552–558.

[5] K. Elvingson, A. González Baró and L. Pettersson, *Inorg. Chem.*, 1996, **35**, 3388–3393.

[6] N. Etxebarria, L. A. Fernández and J. M. Madariaga, *J. Chem. Soc., Dalton Trans.*, 1994, 3055–3059.

[7] G. J.-P. Deblonde, A. Moncomble, G. Cote, S. Bélair and A. Chagnes, *RSC Adv.*, 2015, **5**, 7619–7627.

[8] L. Pettersson, I. Andersson and L. O. Oehman, *Inorg. Chem.*, 1986, **25**, 4726–4733.

[9] L. Pettersson, B. Carlsson, S. Rundqvist, A. F. Andresen and P. Fischer, *Acta Chem. Scand.*, 1975, **29a**, 677–689.

[10] E. Petrus, M. Segado-Centellas and C. Bo, *Inorg. Chem.*, 2022, **61**, 13708–13718.

[11] J. Buils, D. Garay-Ruiz, M. Segado-Centellas, E. Petrus and C. Bo, *Chem. Sci.*, 2024, **15**, 14218–14227.