Supplementary Information for:

Computational Insights into Aqueous Speciation of Metal-Oxide NanoClusters: An In-Depth Study of the Keggin Phosphomolybdate

Jordi Buils,[1,2] Diego Garay-Ruiz,[1,*] Mireia Segado-Centellas,[1,2] Enric Petrus,[1,3] Carles Bo [1,2,*]

[1]Institute of Chemical Research of Catalonia (ICIQ), The Barcelona Institute of Science and Technology, Av. Països Catalans 16, 43007 Tarragona (Spain)
[2]Departament de Química Física i Química Inorgànica, Universitat Rovira i Virgili (URV), C/Marcel·lí Domingo, 43007 Tarragona (Spain)
[3]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, (Switzerland)

Table of contents

Scaling of DFT Formation Constants

POMSimulator calculates the formation constants of the species in the molecular set for every speciation model. Nonetheless, these values were found to be overestimated respect to the experimental constants[1–3] hence the need for a scaling. Originally, the DFT-based formation constants were scaled with a linear regression, comparing the calculated values with the experimental ones found in literature. Then, the speciation models were sorted according to the RMSE value in order to determine the most accurate speciation model, reaction mechanisms and phase diagrams.

However, experimental values are scarce in some systems, and that leads to incomplete regressions with few data points. From previous work, we observed that there seemed to be a *universal scaling* (a unique set of slope and intercept values) for the DFT formation constants, which would allow us to tackle these more sparsely characterized POM systems. However, the use of these unique scaling parameters would break down the current assessment of speciation model quality based on the RMSE against experimental parameters, as this regression will not be carried out anymore. We further discuss this issue in the "Theoretical Background and New Developments" section of the main text.

Because experimental formation constants for PMo Keggin system were available in the literature,[4–6] we employed them to scale the DFT formation constants in order to further investigate the universality of the methodology. Nevertheless, instead of choosing the "best model", we used the average scaling parameters to the totality of the speciation models.

To apply this new approach to the PMo-Keggin system, we needed to first comprehend the overall behavior of all speciation models towards the linear scaling. We generated two density plots (Figure S1) for the slope and intercept of the speciation models after applying the linear regression.

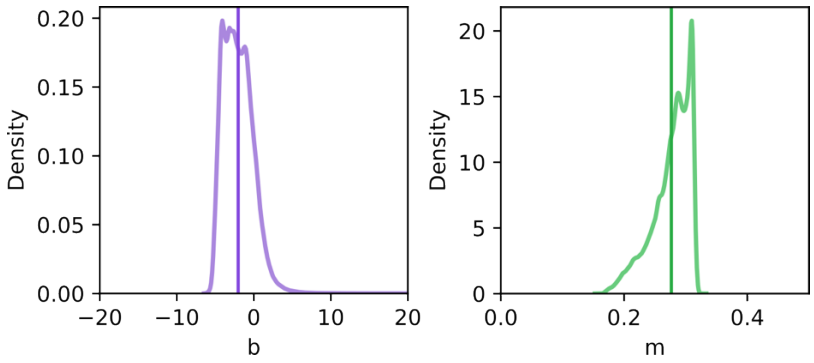

Figure S1. Density plot for the intercept (left) and slope (right) scaling parameters of the PMo system. Vertical lines correspond to the average intercept (-2.02) and slope (0.28) values.

With these density plots we saw that on average the distributions for the slope and intercept parameters were not erratic, and therefore, they could potentially work for all the models. For this reason, we computed the speciation using a unique slope and intercept of 0.28 and -2.02, respectively.

Methodology validation

To validate the statistics-based workflow explained in the main text, we used the previously studied IPA systems for W, Nb and V, in order to assess how the new approach coped with the speciation of the three polyoxometalate systems.

Tungsten

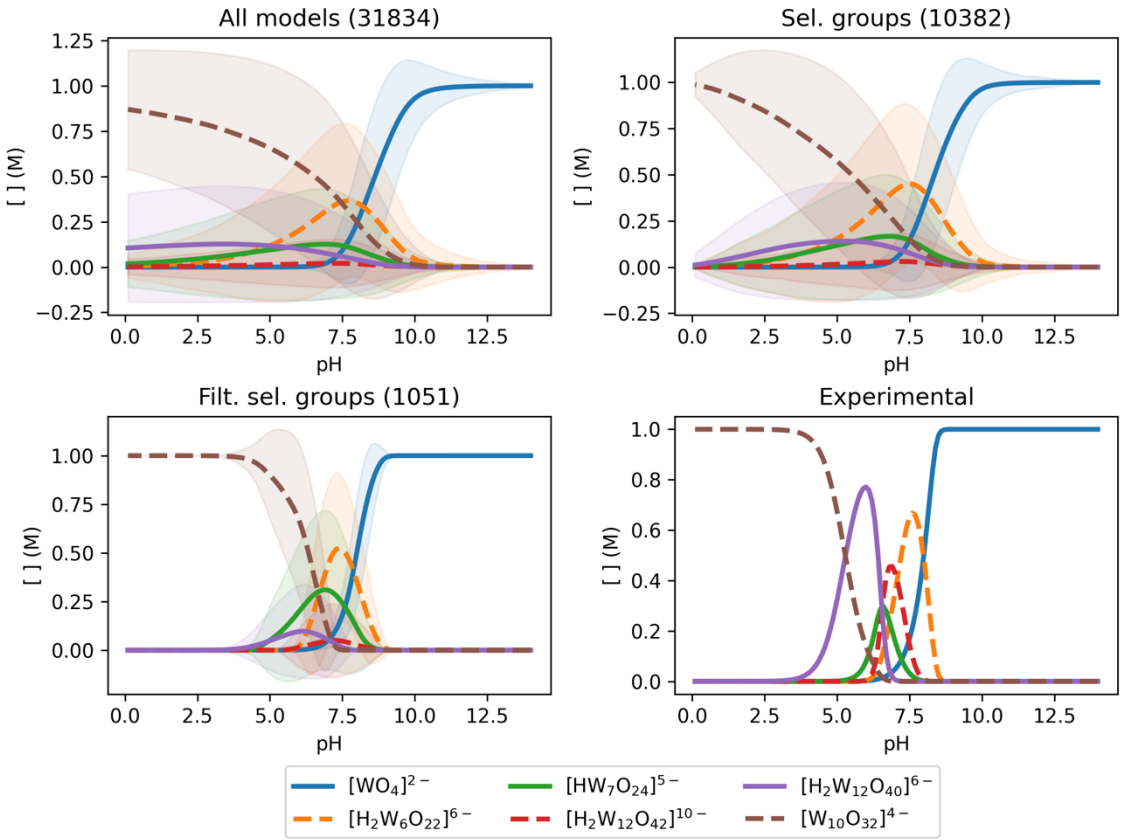

Figure S2. Speciation diagrams for W-IPA systems, including average molar fractions for key species in the chemistry of the system. Top left: diagram with all 31k valid speciation models. Top right: selection of K-Means based clusters in agreement with experimental observations. Bottom left: remaining models after discarding outliers on the speciation of $W_{10}O_{32}$-0H, Bottom right: simulated diagram from reported experimental constants[7].

In Figure S2 we can observe how the statistical treatment pipeline refines the average speciation diagram until reaching a reasonable match compared to the simulation using the experimental constants (bottom right). While the average of the 31k models (top left) and the selected cluster (top right) are quite similar, there are some relevant modifications. For example, the increase in the average W_{10} cluster dominant at acidic pH. Moreover, filtering by this same W_{10} species produces the plot in the bottom left corner, where peak positions are much more defined and in good agreement with the expected results, and all relevant species are represented. While of course there is not a perfect agreement between the clustered-filtered average speciation and the experimental results (e.g., peaks for the W_{12} species are lower than expected). we believe that the generality of the approach and its scalability to larger and more fuzzily described systems compensates for this loss in accuracy.

The eight groups produced by the clustering are shown in Figure S3. From these, we chose **Cluster 1** as the most representative of the speciation of tungsten, as it was the group in which the $\{W_6\}$, $\{W_7\}$ and $\{W_{10}\}$ metal-oxo clusters, observed in the experiments, were most represented.

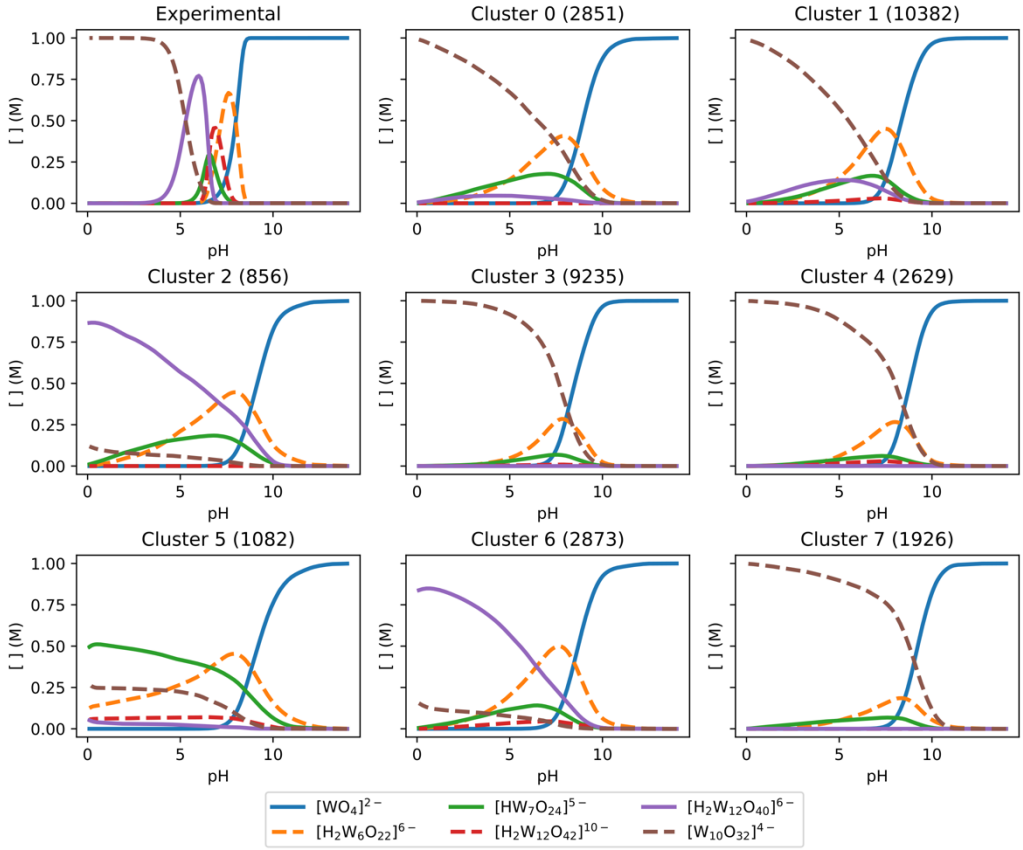

Figure S3. Average speciation diagram for the eight clusters generated for the W-IPA system. Experimental diagram (top left) is included as reference.

Additionally, we explored how the group distribution reflected the RMSE of each model (Figure S4). Here, we clearly observe how the models with lower RMSEs, which were the ones selected with the original methodology of POMSimulator, mainly fall in **Cluster 1**. Therefore, our chemistry-oriented selection of groups does agree with the purely mathematical RMSE-based approach.

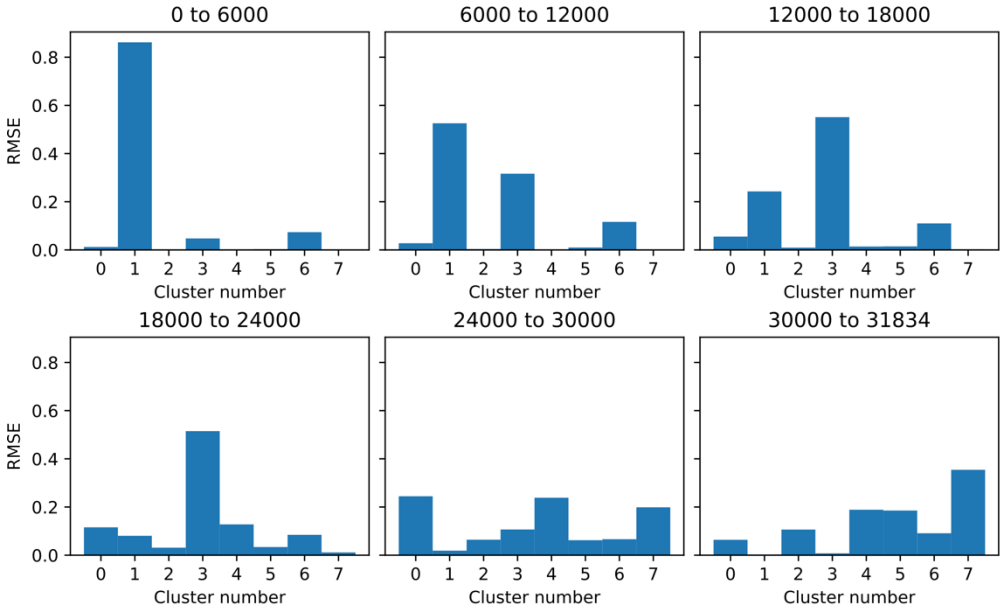

Figure S4. Comparison of RMSE-based and cluster-based model groupings. Each panel corresponds to the histogram of the cluster label in each range of models ordered by RMSE, from most accurate models (0 to 6000, top left) to worst (bottom right).

Niobium

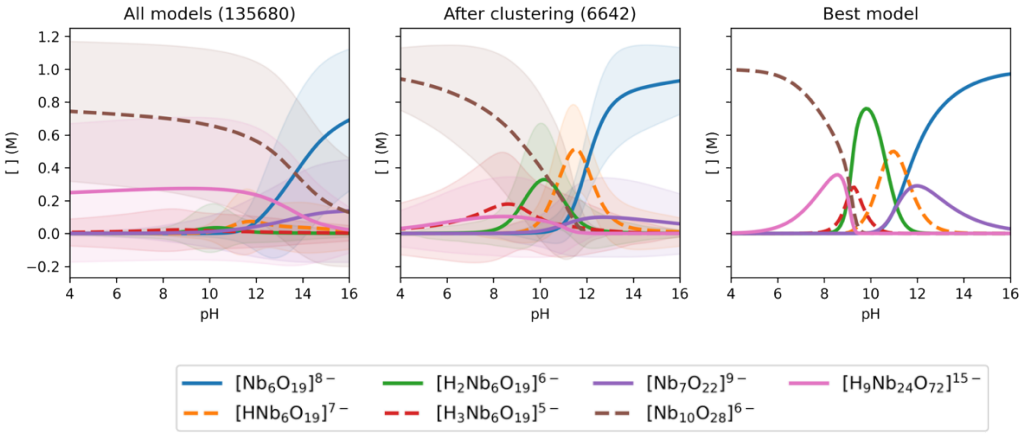

Figure S5. Speciation diagrams for Nb-IPA systems, including average molar fractions for key species in the chemistry of the system. Left: diagrams with all 136k valid speciation models. Center: selection of K-Means based clusters in agreement with experimental observations. Right: speciation diagram based on the best speciation model characterized by POMSimulator.

Given that Nb had more speciation models than W, we applied *two* clustering steps. In the first clustering, we generated eleven groups and discarded the ones that did not match the main species at the limits of the diagram: that is, Nb_{10} at the acidic side and Nb_6 at the alkaline one. For the second clustering, we considered eight groups and selected the one with the best representation of the rest of the relevant species in the system, leading to the speciation diagram in the center of Figure S5. In this case, the second step of the workflow (removal of outliers for a target species) did not improve the agreement with experimental results , and thus, we selected the full cluster instead.

Comparing the three speciation diagrams, we observe that the average of all 135k models (left in Figure S5) is not only inferior from the speciation point of view [1], but also it depicts high variances for almost all species. In stark contrast, the diagram obtained after clustering (center of Figure S5) is remarkably close to this best model, validating the robustness of the statistical treatment.

At this point and recalling our limitations on the characterization of our target system (PMo-HPAs), we decided to test the performance of the current pipeline on a subset of the Nb system. We sampled 10% of the entries in the dataset, arriving at a small set of 13.5k models. Next, we applied the same pipeline as in the whole set, obtaining the speciation diagrams in Figure S6. From there, we observe an excellent agreement, with the all-model averages being extremely similar between the complete population and the sample. Moreover, the selected group of models has the same qualitative behavior as the previous one, with even a slight improvement in the description of the triprotonated Nb_6 species. This agreement pinpoints the adequacy of the random sampling we performed among the PMo-HPA model set.

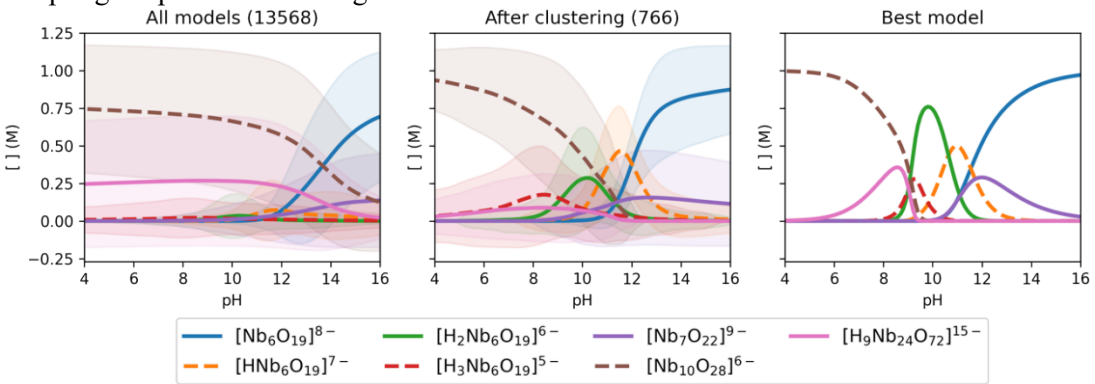

Figure S6. Speciation diagrams for a random subsample of 10% of the Nb-IPA systems, including average molar fractions for key species in the chemistry of the system. Left: diagram with the complete 13.5k valid speciation models.

Center: diagram with the selection of K-Means based clusters in agreement with experimental observations. Right: speciation diagram based on the best model characterized in our previous work[1].

Vanadium

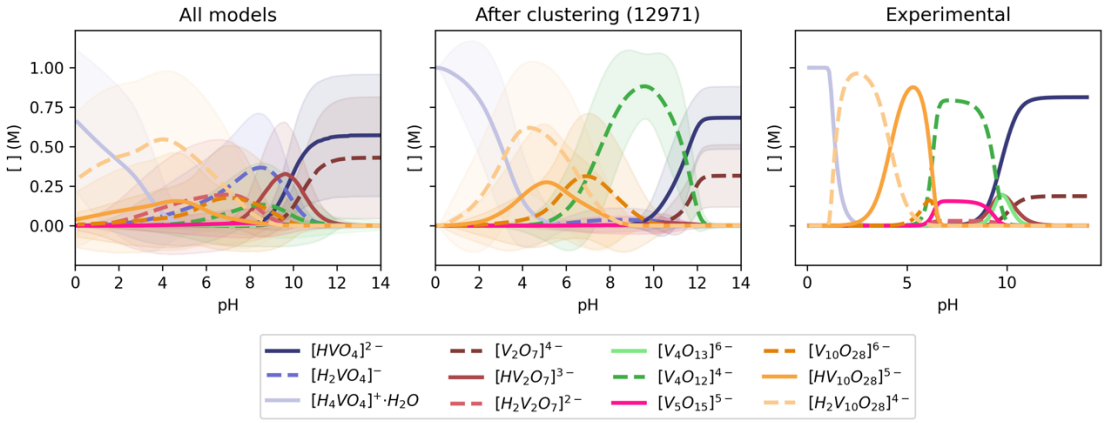

Figure S7. Speciation diagrams for V-IPA systems, including average molar fractions for key species in the chemistry of the system. Left: diagrams with 855k valid speciation models. Center: diagram with the selection of K-Means based clusters in agreement with experimental observations. Right: simulated diagram from the reported experimental constants by Elvingson et al.[8]

The V-IPA system shows two major differences with the previous examples (Nb and W). Not only does it include more speciation models (up to 1 million, even though reduced to 855k after removing not converged models), but also the target speciation diagram is more complex. This is due to the amphoteric nature of vanadium: while the behavior of W is acidic and that of Nb is alkaline, the assembly of vanadates involved a larger pH range (as we reported in previous studies[1]).

Following the guidelines for niobium, we applied two clustering stages, both with 11 groups, selecting the diagrams that were closer to the expected speciation. Again, filtering outliers for selected species in the system did not improve the final agreement with the target speciation diagram, and thus Figure S7 does not take this part of the protocol into account.

Despite the complexity of the experimental speciation diagram, we achieved a reasonable agreement. While not all species are represented, like the $\{V_5\}$ cluster, most other important features are reflected: the switch between hydrated $H_4VO_4^+ \cdot H_2O$ and $\{V_{10}\}$ at acidic pH, the coexistence of monomers and dimers at the alkaline limit, and the formation of $\{V_4\}$ metavanadates between neutral and alkaline regions.

Overall, the success in the representation of acidic (W), alkaline (Nb) and amphoteric (V) behaviors through our clustering pipeline, using the same average slope and intercept parameters for the whole datasets, and with vastly different numbers of speciation models, confirms the robustness of the approach. While the agreement is not as accurate as with the lowest-RMSE model strategy, the inherent generality of the clustering approach provides a steppingstone to use POMSimulator for more complex systems, such as the phosphomolybdates treated in this work.

Feature subset exploration

We have proposed four descriptors for the featurization of speciation diagrams (see Figure S8): position, height, width and area of each peak. To assess how many of these are strictly necessary, we considered the tungsten system and applied the clustering strategy with each possible combination where these features are or not present, for a total of 15 different possibilities. Here, we applied the clustering and selected the cluster that was more similar to the expected speciation diagram.

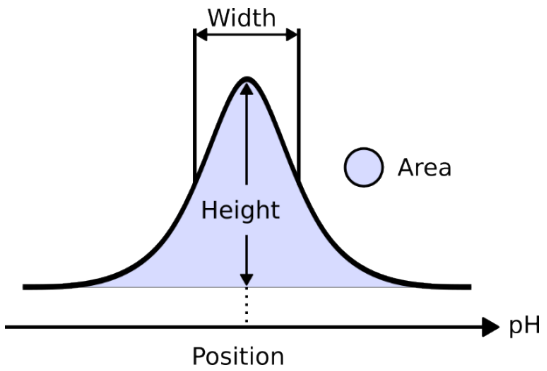

Figure S8 Graphical description of selected features for the speciation peaks. For each SM, we will select the concentration curve of each molecule and extract four features: 1) the peak position which is the pH value with the highest concentration, 2) the peak height as the maximum concentration value, 3) the peak width calculated as the width of the concentration curve at the height value divided by two, and last, 4) the peak area.

As shown in Figure , the choice of descriptors is not trivial. While there are sets that actually give considerably inaccurate predictions (e.g. area, width+area, width+height+area), most of the others are rather similar. Considering the limited success of the area in this dataset, we ended up selecting the combination **width + position + height**. It does not only provide the best-defined peaks for all species at intermediate pH values, but also it conceptually covers all topological aspects. Nonetheless, depending on the system, it could be argued that if this descriptor set does not provide a good separation between groups, other combinations might be tested without compromising the application of the methodology.

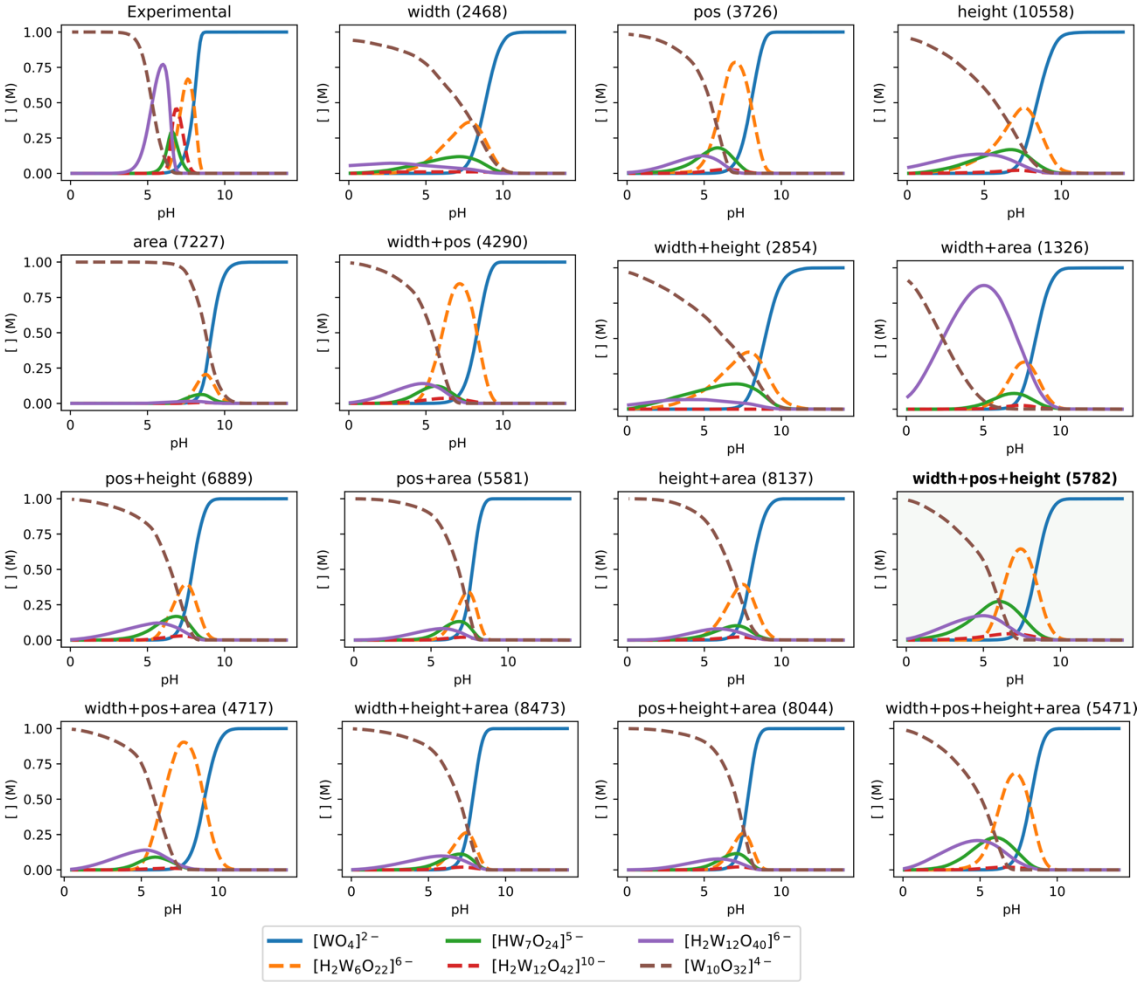

Figure S9. Exploration of feature subsets for the W-IPA system. For every combination we show the number of models in the selected group between parentheses. The selected feature combination (width + pos + height) is highlighted in bold.

Evaluating the effect of the seed in random sampling

As depicted in Figure S6, even small random samples provide a satisfactory agreement with the speciation diagrams produced from the complete population. For the sake of completeness, we considered different random samples over the Nb dataset and applied the first stage of the clustering protocol, selecting a single group as the most representative of the expected speciation diagram. Here, we did not proceed with the second clustering, as we only strived to identify the similarity between the results of the complete population and the samples.

Figure S10 shows the robustness of the clustering approach: the average speciation predicted through the selected group of models is consistent between the eight random seeds tested and the original population. While there is a certain variability, with the groups in the lower left and lower center of Figure S10 having a smaller number of models and a greater peak for the $\{Nb_{24}\}$ cluster, the approach is still robust enough, validating the strategy followed for the phosphomolybdate system characterized in the main text.

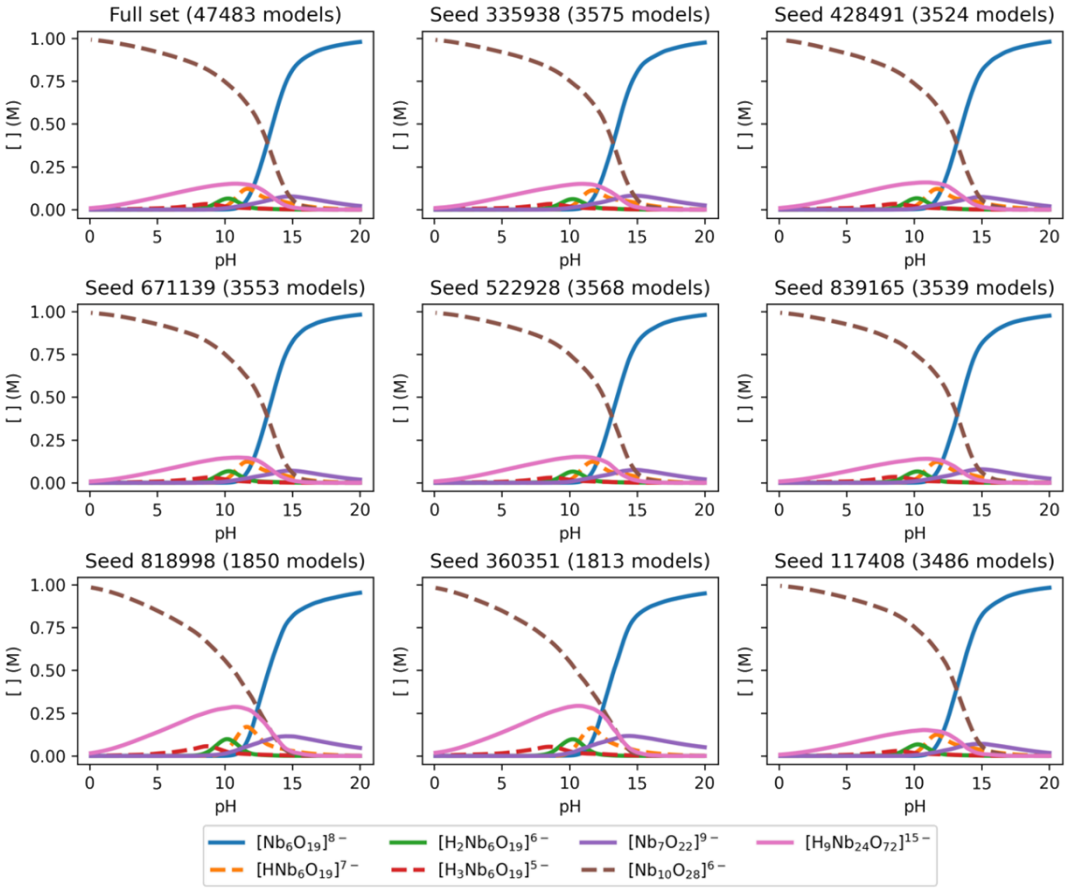

Figure S10. Exploration of different initial seeds for the random sampling of speciation models on the Nb-IPA dataset.

Comparison of the two approaches for selecting SMs

In Figure S11 we depict a comparison of the new (clustering-based) and old (RMSE-based) speciation methodologies. As demonstrated in previous studies, the speciation for IPAs using a single model was consistent with experimental data. On the other hand, the high complexity HPAs systems strongly limits the accuracy of the prediction. The lowest RMSE models did not match the experimental results. Only a few models among the 500 lowest RMSE models, did agree, at least partially, with the results from Cadot et al.[4] In contrast, as explained in the main text, a clustering-based methodology matches the experimental speciation diagram more closely.

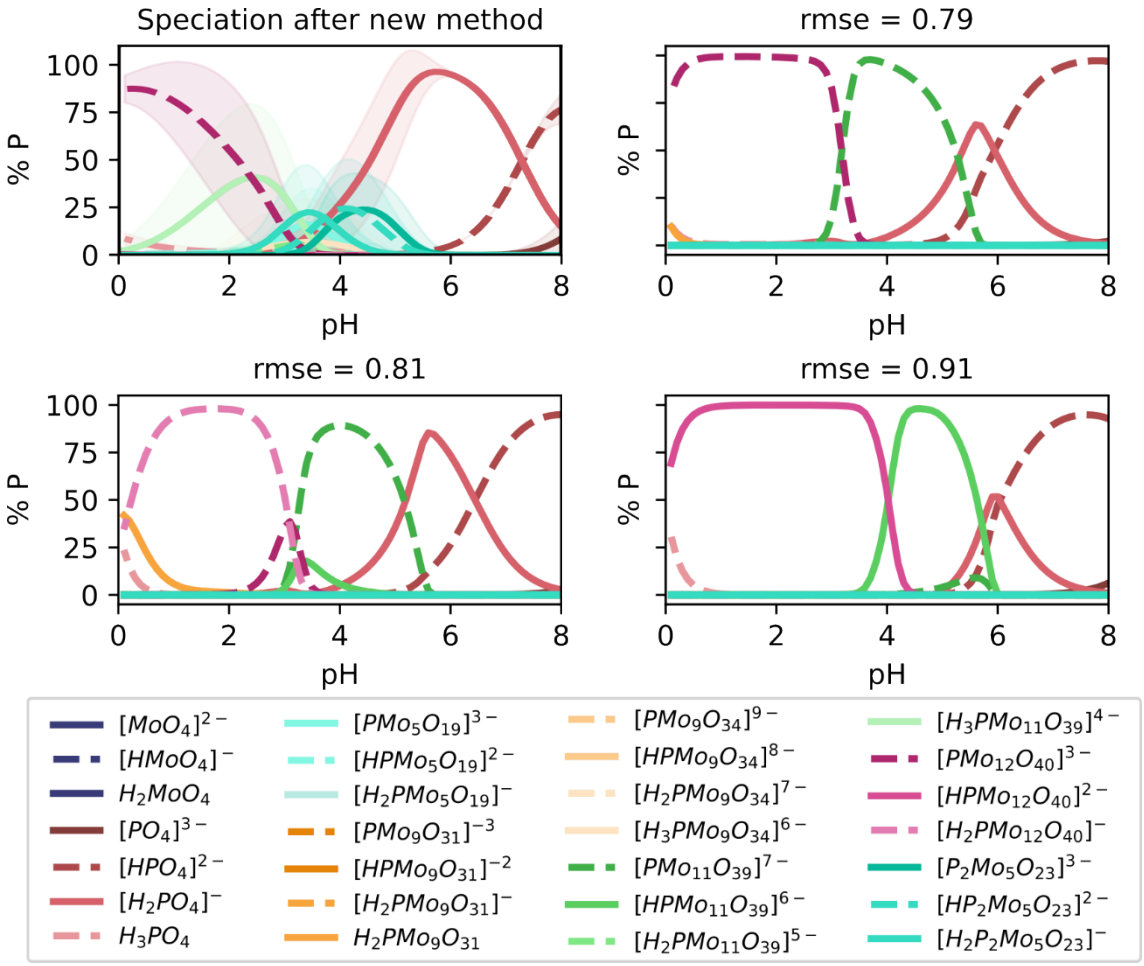

Figure S11 Comparison of new (top left pane) and old (top right and bottom panes) methodology for speciation diagrams.

Model selection for phase diagram generation

As discussed throughout the main text, the selection of SM clusters is based on the chemistry-informed analysis of their corresponding average speciation diagrams. In the case of the PMo system, all the computations leading to the final SM selection ($2.5 \cdot 10^4$ models) considered a ratio [Mo]/[P] = 12, in line with experimental results. However, in order to properly compute the phase diagram throughout varying ratios ($1 \leq [Mo]/[P] \leq 12$), we decided to consider a more diverse set of models. In this way, more chemically reasonable models could be taken into account, as they might become important at these lower molybdenum/phosphorus ratios. Consequently, we added additional groups of SMs from the very same clustering employed for the speciation, accounting for a total of 76917 SMs. It is worth noting that these additional models were qualitatively quite similar to the models used for the speciation diagram in Figure 3, with analogous species and peak positions, mainly differing on the abundance of some of the species. Therefore, these SMs

can be regarded as reasonable descriptions of the system, and thus possibly more relevant when varying conditions like the Mo/P ratio. On the other hand, they might also be safely neglected when more detailed experimental information is available, as in the case of the well-defined experimental speciation diagrams reported by Cadot et al.[4]

Formation constants for phosphorus-molybdate system

Table S1. Scaled DFT formation constants for the 42 species in the PMo_{12} system. Scaling expression is y=0.28x-2.02. Formation reactions defined respect to $[MoO_4]^{2-}$ and $[PO_4]^{3-}$.

Polyoxometalate	Mean ± Standard Deviation
$[MoO_4]^{2-}$	-2.02 ± 0.00
$[HMoO_4]^{-}$	4.28 ± 0.00
H_2MoO_4	6.15 ± 0.01
$[Mo_2O_7]^{2-}$	10.77 ± 0.04
$[HMo_2O_7]^{-}$	13.33 ± 0.04
$H_2Mo_2O_7$	14.82 ± 0.06
$[HMo_2O_8]^{3-}$	-1.53 ± 0.12
$[H_2Mo_2O_8]^{2-}$	7.66 ± 0.18
Mo_3O_9	22.35 ± 0.16
$[HMo_3O_9]^{+}$	17.86 ± 0.84
$[H_2Mo_3O_9]^{2+}$	16.01 ± 1.76
$[Mo_3O_{10}]^{2-}$	19.80 ± 0.15
$[HMo_3O_{10}]^{-}$	21.47 ± 0.16
$H_2Mo_3O_{10}$	22.35 ± 0.18
$[HMo_3O_{11}]^{3-}$	13.81 ± 0.98
$[H_2Mo_3O_{11}]^{2-}$	19.16 ± 1.05
$[HMo_6O_{21}]^{5-}$	34.47 ± 1.44
$[H_2Mo_6O_{21}]^{4-}$	37.86 ± 1.93
$[PO_4]^{3-}$	-2.02 ± 0.00
$[HPO_4]^{2-}$	9.07 ± 0.37
$[H_2PO_4]^{-}$	16.37 ± 0.37
H_3PO_4	19.16 ± 0.37
$[PMo_3O_{13}]^{3-}$	30.83 ± 0.94
$[HPMo_3O_{13}]^{2-}$	38.25 ± 0.94
$[H_2PMo_3O_{13}]^{-}$	42.31 ± 0.95
$H_3PMo_3O_{13}$	42.92 ± 0.95
$[PMo_5O_{19}]^{3-}$	58.24 ± 1.00
$[HPMo_5O_{19}]^{2-}$	60.75 ± 1.00
$[H_2PMo_5O_{19}]^{-}$	60.10 ± 1.02
$[PMo_6O_{22}]^{3-}$	67.33 ± 1.12
$[HPMo_6O_{22}]^{2-}$	69.48 ± 1.13
$[PMo_9O_{31}]^{3-}$	91.78 ± 2.18
$[HPMo_9O_{31}]^{2-}$	90.38 ± 2.50
$[H_2PMo_9O_{31}]^{-}$	91.14 ± 2.57
$H_3PMo_9O_{31}$	91.08 ± 2.65
$[PMo_9O_{34}]^{9-}$	72.96 ± 2.20
$[HPMo_9O_{34}]^{8-}$	76.29 ± 1.99
$[H_2PMo_9O_{34}]^{7-}$	80.05 ± 1.93
$[H_3PMo_9O_{34}]^{6-}$	83.88 ± 2.89

$[PMo_{11}O_{39}]^{7-}$	96.94 ± 2.19
$[HPMo_{11}O_{39}]^{6-}$	103.02 ± 2.41
$[H_2PMo_{11}O_{39}]^{5-}$	108.38 ± 2.62
$[H_3PMo_{11}O_{39}]^{4-}$	112.19 ± 2.64
$[PMo_{12}O_{40}]^{3-}$	124.55 ± 2.57
$[HPMo_{12}O_{40}]^{2-}$	122.09 ± 2.45
$[H_2PMo_{12}O_{40}]^{-}$	118.96 ± 2.48
$[P_2Mo_5O_{23}]^{6-}$	69.99 ± 1.59
$[HP_2Mo_5O_{23}]^{5-}$	73.04 ± 2.80
$[H_2P_2Mo_5O_{23}]^{4-}$	76.42 ± 3.23

References

(1) Petrus, E.; Segado-Centellas, M.; Bo, C. Computational Prediction of Speciation Diagrams and Nucleation Mechanisms: Molecular Vanadium, Niobium, and Tantalum Oxide Nanoclusters in Solution. *Inorganic Chemistry* **2022**, *61* (35), 13708–13718. https://doi.org/10.1021/acs.inorgchem.2c00925.

(2) Petrus, E.; Segado, M.; Bo, C. Nucleation Mechanisms and Speciation of Metal Oxide Clusters. *Chemical Science* **2020**, *11* (32), 8448–8456. https://doi.org/10.1039/d0sc03530k.

(3) Petrus, E.; Bo, C. Unlocking Phase Diagrams for Molybdenum and Tungsten Nanoclusters and Prediction of Their Formation Constants. *Journal of Physical Chemistry A* **2021**, *125* (23), 5212–5219. https://doi.org/10.1021/acs.jpca.1c03292.

(4) Yao, S.; Falaise, C.; Leclerc, N.; Roch-Marchal, C.; Haouas, M.; Cadot, E. Improvement of the Hydrolytic Stability of the Keggin Molybdo- and Tungsto-Phosphate Anions by Cyclodextrins. *Inorganic Chemistry* **2022**, *61* (9), 4193–4203. https://doi.org/10.1021/acs.inorgchem.2c00095.

(5) Gumerova, N. I.; Rompel, A. Polyoxometalates in Solution: Speciation under Spotlight. *Chemical Society Reviews* **2020**, *49* (21), 7568–7601. https://doi.org/10.1039/d0cs00392a.

(6) Pettersson, L.; Andersson, I.; Óhman, L.-O. Contribution from the Speciation in the Aqueous H+-Mo042-HP042 System As Deduced from a Combined Emf-31P NMR Study*. *Inorg. Chem* **1986**, *25*, 4726–4733. https://doi.org/10.1021/ic00246a028.

(7) Rozantsev, G. M.; Sazonova, O. I. Thermodynamic Parameters of Interconversions of Isopolyanions in Solutions of Tungsten(VI). *Russ. J. Coord. Chem.* **2005**, *31* (8), 552–558. https://doi.org/10.1007/s11173-005-0135-x.

(8) Elvingson, K.; González Baró, A.; Pettersson, L. Speciation in Vanadium Bioinorganic Systems. 2. An NMR, ESR, and Potentiometric Study of the Aqueous H^+ −Vanadate−Maltol System. *Inorg. Chem.* **1996**, *35* (11), 3388–3393. https://doi.org/10.1021/ic951195s.