Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach

Jordi Buils Casasnovas

UNIVERSITAT ROVIRA i VIRGILI

ICIQ Institut Català d'Investigació Química

Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach

Jordi Builds Casasnovas

DOCTORAL THESIS
2024

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Jordi Builds Casasnovas

Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach

Doctoral Thesis

Supervised by Prof. Carles Bo Jané and Dr. Mireia Segado Centellas

Tarragona

2024

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Prof. Dr. Carles Bo Jané, group leader at the Institute of Chemical Research of Catalonia (ICIQ) and professor at Universitat Rovira i Virgili (URV) and Dr. Mireia Segado Centellas, researcher at Universitat Rovira i Virgili:

WE STATE that the present study, entitled "Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach", presented by Jordi Builds Casasnovas for the award of the degree of Doctor, has been carried out under our supervision at the Institute of Chemical Research of Catalonia.

Tarragona, October 2024

Prof. Carles Bo Jané

Carles Bo Jané
1 Oct 2024 12:46:02 CEST
Certificat emès per: EC-SectorPublic
Número de serie: 14395090865545649814502245244

Dr. Mireia Segado Centellas,

Firmado por SEGADO CENTELLAS, MIREIA (FIRMA) el día 01/10/2024 con un

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Acknowledgements

I would like to thank my thesis supervisors Prof. Carles Bo and Dr. Mireia Segado-Centellas for giving me the opportunity to work in the group, and for the help throughout the past 3 years. I would also like to thank my group members for their unvaluable support and help.

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Funding agencies

The work presented in this PhD thesis has been funded by the Institute of Chemical Research of Catalonia (ICIQ) and by FPI grant (PID2020-112806RB-I00/MICIU/AEI/10.13039/501100011033) from the Spanish of Science, Innovation and Universities.

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Publications

Enric Petrus, Jordi Builds, Diego Garay-Ruiz, Mireia Segado-Centellas, and Carles Bo. POMSimulator: An Open-source Tool for Predicting the Aqueous Speciation and Self–Assembly Mechanisms of Polyoxometalates. *Journal of Computational Chemistry* 45, 26: 2242-50. https://doi.org/10.1002/jcc.27389.

Jordi Builds, Diego Garay-Ruiz, Mireia Segado-Centellas, Enric Petrus, and Carles Bo. Computational Insights into Aqueous Speciation of Metal-Oxide Nanoclusters: An in-Depth Study of the Keggin Phosphomolybdate. *Chemical Science*, 2024. https://doi.org/10.1039/D4SC03282A.

Jordi Builds, Diego Garay-Ruiz, Enric Petrus, Mireia Segado-Centellas, Carles Bo. Towards a Universal Scaling Method for Predicting Equilibrium Constants of Polyoxometalates. *ChemRxiv*, 2024. https://doi.org/10.26434/chemrxiv-2024-r2lsq

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Abstract

In this thesis we have expanded the applicability of POMSimulator from isopolyoxometalates to heteropolyoxometalates. To tackle the increasing complexity derived from the expansion to heteropolyoxometalates (HPA) systems, we have developed two main methodologies. First, we propose a general equation for the scaling of formation constants, universal for all polyoxometalates and independent of the Density Functional Theory (DFT) method, based on Multi-Linear Regression (MLR) models. Second, we have developed a statistical workflow based on clustering techniques to manage the large number of speciation models generated by POMSimulator. Using these data-driven approaches we have successfully simulated the aqueous speciation of two HPA systems such as the phosphomolybdate (PMo) and the arsenomolybdate (AsMo) in agreement with experimental results. We have predicted the formation of the well-known PMo Keggin $\{PMo_{12}\}$ anion in contrast with the absence of the equivalent structure in the AsMo system. We have also confirmed the importance of the $\{As_2Mo_6\}$ and $\{AsMo_9\}$ species in the arsenomolybdate system. We have reported the first speciation phase diagrams for HPA systems giving a general overview on the chemistry of heteropolyoxometalates. The current development of

---

POMSimulator opens the door for further studies in the field and builds up the synergy with experimental studies to improve the understanding of the complex self-assembly of polyoxometalates.

Contents

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Chapter 1

Introduction to polyoxometalates

1.1 Polyoxometalates

Polyoxometalates (POM) are molecular anionic metal-oxo clusters, consisting of two or more metal atoms, usually from groups V (V, Nb, Ta) and VI (Mo, W) in the highest oxidation state, combined through metal-oxygen-metal motifs. Figure 1.1 represents some examples of polyoxometalates.

The first ever report on a POM is dated back to 1826 by J.J. Berzelius[1], in which the first ammonium phosphomolybdate ($[NH_4]_3[PMo_{12}O_{40}]$) was introduced. But it wasn't until 1933 when J. F. Keggin characterized it by X-ray crystallography[2]. Many years later, with the introduction of new synthetic methods and the introduction of ^{17}O nuclear magnetic resonance (NMR) spectroscopy, the characterization of the Keggin anion in solution was possible. In 1926 the Wells-Dawson structure $[X_2M_{18}O_{62}]^{6-}$

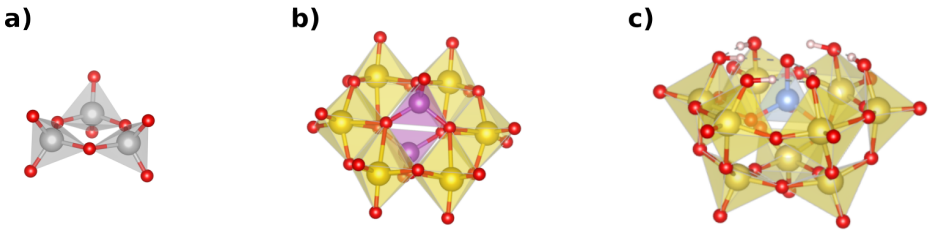

Figure 1.1: Various polyoxometalates structrues: a) W_3O_9 Building Block, b) As_2Mo_6 and c) $H_6SiMo_9O_{34}$. Red spheres correspond to oxygen atoms, grey atoms and polyhedra correspond to tungsten atoms, yellow spheres and polyhedra correspond to molybdenum atoms, purple spheres and polyhedra correspond to arsenic atoms and blue spheres and polyhedra correspond to silicon atoms.

was synthesized for the first time. Later, in 1937 J.S. Anderson proposed a structure for the $[XM_6O_{24}]^{6-}$ motif[3], but it wasn't until 1948 when H. T. Evans determined the structure that Anderson had proposed years before[4], receiving the name after the two scientists. In 1953 the Wells-Dawson structure was characterized by powder X-ray diffraction (XRD)[5]. In the same year I. Lindqvist proposed the structure for the $[M_6O_{19}]^{8-6}$. In 1973 R. Strandberg determined the structure of $Na_6[P_2Mo_5O_{23}] \cdot (H2O)_{13}$ by 3D-XRD[7]. After that in in 1983 M.T. Pope published the book entitled "Heteropoly and Isopoly Oxometalates"[8] which provided a thorough review of the POM characterization done until then. Later, in 1995 and 1998, A. Müller reported the molybdenum wheels Mo_{154}[9] and Mo_{132}[10] respectively. In Figure 1.2 a chronological axis with some of the most important milestones in the field of polyoxometalates is depicted in addition to a plot with the amount of published articles related to POMs across time. From this plot we see that after M.T. Pope's comprehensive work in 1983, which organized previous research efforts, and A. Müller's contributions in 1995 and 1998, particularly with the introduction of giant POMs, the field of polyoxometalates experienced significant growth.

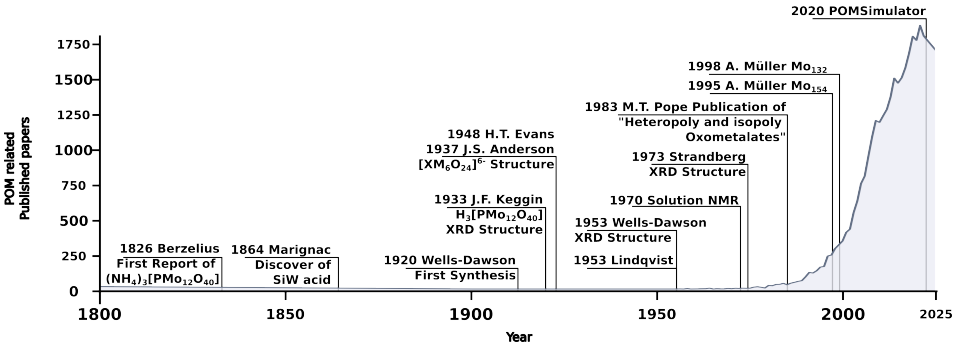

Figure 1.2: Chronological axis of the most important milestones in the history of Polyoxometalates. The number of published papers related to polyoxometalates is shown as a line plot, with its y axis in the left side. The x axis corresponds to the time scale. (Query data for the line plot:[11]).

The synthesis of polyoxometalates depends mostly on three parameters: temperature, ionic strength and pH[12–15]. For example, POMs that contain Mo or W have an acidic chemistry and in high pH values only the monomeric species can be found. Vanadium polyoxometalates have an amphoteric chemistry and species can be found in both acidic conditions and neutral-basic conditions. On the other hand, niobium and tantalum POMs have a basic chemistry, and their speciation is rich in high pH values[14].
Over the past few decades, POMs have been used in a wide range of applications across different fields. For instance, polyoxomolybdates have been used as drugs for cancer treatment[16–18], they have also been used as antiviral drugs[19]. Additionally, POMs can adopt different oxidation states without significant structural changes, maintaining their high stability under harsh conditions. For this reason, POMs are used as photo[20,21] and electro catalysts, making use of their reversible redox properties and chemical stability[22,23]. Furthermore, the ability to accepting electrons has been crucial for synthesizing new materials with electron storage capabilities;

for instance the phosphotungstate Wells-Dawson can reversibly store up to 18 electrons[24]. Polyoxometalates can be classified in heteropolyoxoanions (HPAs) and isopolyoxoanions (IPAs) according to the presence or absence of heteroatoms respectively.

1.1.1 Iso-Polyoxometalates

Isopolyoxometalates or IPAs are molecular polyoxometalates that only contain a metal, oxygen and hydrogen. In some cases, a mixture of different metals can be found: for example, molybdenum and tungsten, tungsten and tantalum[25], niobium and tantalum[26] or tungsten and vanadium[27]... From this family of molecules, there are some structures that stand out over the rest, such as the Lindqvist, the decametalate or the Mo_{132} Keplerate. Their synthesis is quite straight forward, dissolving an inorganic salt in water, adjusting pH and ionic strength, temperature and incubation time[15]. In some cases, reducing agents are also included in the reaction, giving place to mixtures of metals in different oxidation states. An example of this can be found in the Mo_{132} Keplerate, where 2 different motifs are present. The first one is a $Mo_2O_2X_2$ (X=O, S, Se...) $\{Mo_2\}$ which contains two Mo(V), featuring a metal-metal bond and used as a linker. The second is a star-shaped Mo_6O_{21} $\{Mo_6\}$ which contains 6 Mo(VI). The aggregation of 30 Mo_2 linkers and 12 Mo_6 stars forms the Mo_{132} $[Mo(V)_{60}Mo(VI)_{72}]$. In Figure 1.3 4 examples of isopolyoxometalates are depicted. From left to right: on the first place we find the Nb Lindqvist that corresponds to $[Nb_6O_{19}]^{8-}$.

In second place the V decametalate is shown, which corresponds to $[V_{10}O_{28}]^{8-}$. The decametalate is similar to the Lindqvist anion, as if two molecules had merged. In its structure, it is possible to find the structure of

the Lindqvist, sharing a corner with 4 other metals. Next, the Nb_{24} is depicted, which corresponds to $[H_9Nb_{24}O_{72}]^{15-}$. This structure is formed by three units of Nb_7 joined by 3 $[H_2NbO_5]$ units, forming a trimer structure. Last, the $Mo132$ Keplerate is shown corresponding to $[Mo_{132}O_{372}]^{12-}$. As mentioned before, the structure can be divided between ${Mo^V}_2$ linkers (grey) and ${Mo^{VI}}_6$ units (yellow).

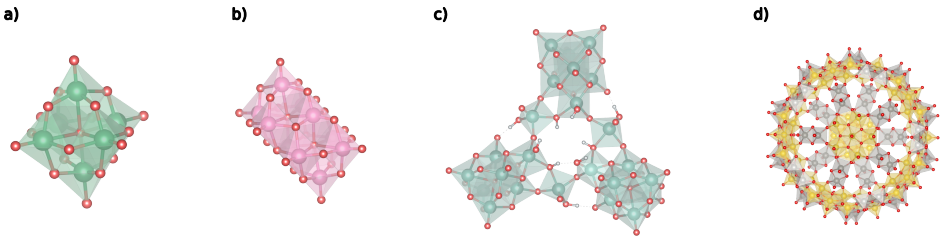

Figure 1.3: Important isopolyoxometalates structures: a) Nb_6 Lindqvist, b) V_{10} Decavanadate, c) Nb_{24} and d) Mo_{132} Keplerate.Red spheres correspond to oxygen atoms, green atoms and polyhedra correspond to niobium atoms, yellow and grey spheres and polyhedra correspond to molybdenum atoms and pink spheres and polyhedra correspond to vanadium atoms.

1.1.2 Hetero-Polyoxometalates

Opposite from IPAs, heteropolyoxometalates contain one or more heteroatom in the center of the structure[8].Most common heteroatoms are phosphorus[28] or silicon[14], but others like arsenic[29], boron[30] or aluminum[31,32] have also been employed. Heteroatoms inside polyoxometalates adopt a tetrahedral geometry in opposition of the metals on the metal-oxo framework that can be found either in octahedral or square pyramidal.

Among the high number of heteropolyoxometalates, there is one that has to be highlighted for its history which is the Keggin structure. First reported by Berzelius, and studied by Marignac and Pauli, it wasn't until 1933 that J.F. Keggin determined its structure by powder XRD[2]. Keggin

structures can be described as $[XM_{12}O_{40}]^{n-}$ where X is usually P or Si, and M is commonly Mo or W although other kind of metals have been used. Other important hetero-POMs are the lacunary POMs, which are polyoxometalates with one or more missing metal-oxygen polyhedral units, offering unique properties and reactivity. These vacant sites can be used to anchor a variety of metal ions, organic molecules or clusters, making lacunary structures versatiles building blocks for creating novel materials. For instance, if we subtract a MO from the $[XM_{12}O_{40}]^{n-}$ we obtain a $\{XM_{11}\}$ structure called Keggin lacunary[33] which can then be used as framework to insert other metal atoms in the vacancy generated to add other properties. If for instance, we subtract 3 metal atoms from the $\{XM_{12}\}$ structure, we obtain $\{XM_9\}$[34] structure that can suffer a dimerization process to form the Wells-Dawson[5] POM which has been employed for its electron storage properties in electrocatalysis[24]. Figure 1.4 represents three well-known heteropolyoxometalates:

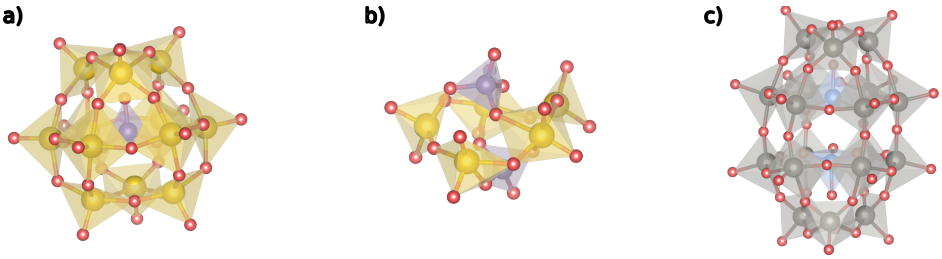

Figure 1.4: Important heteropolyoxometalates structures: a)PMo_{12} Keggin, b) P_2Mo_5 Strandberg and c) Si_2W_{18} Wells-Dawson. Red spheres correspond to oxygen atoms, grey atoms and polyhedra correspond to tungsten atoms, yellow spheres and polyhedra correspond to molybdenum atoms, purple spheres and polyhedra correspond to phosphorus atoms and blue spheres and polyhedra correspond to silicon atoms.

1.2 Experimental speciation studies on polyoxometalates

Polyoxometalates synthesis takes place in aqueous media in almost all cases[15]. This is why, it is important to understand the behaviour of POMs in water solution, and how they interact with other POMs or with the solvent. One of the most accepted ideas is that polyoxometalates form through self-assembly processes[35–38], in which there are building blocks which are initially formed and then grow into bigger clusters.

From the experimental point of view, solution NMR[14] has become quite an important tool to understand POMs structures in solution, and how the different species distribute. For IPAs, ^{17}O has been widely used to discern between different structures, along with ^{183}W or ^{95}Mo NMR and other active nuclei (V, Si, H).

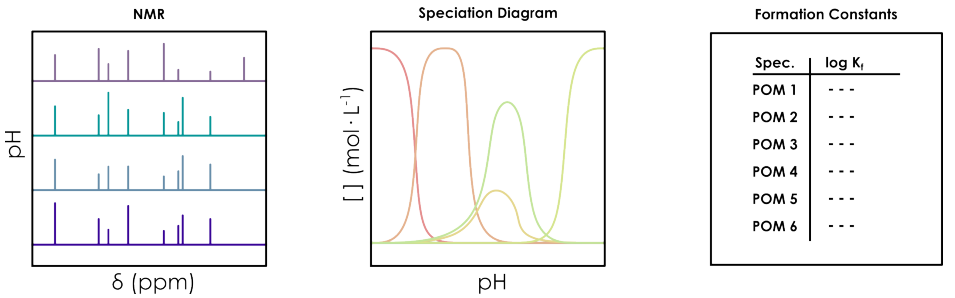

Figure 1.5: Graphical example of the conversion of a NMR spectra to formation constants for polyoxometalates. From left to right, first image represents the NMR spectra acquirement at different pH values. Second image corresponds to the construction of the speciation diagram. Last image corresponds to the calculation of the formation constants.

In phosphorus HPAs, ^{31}P has been widely used, as different POMs have different ^{31}P-NMR fingerprints[28]. This property allows the deconvolution of NMR spectra into the sum of the signals of each of the species present in

solution. Then using a linear combination it is possible to obtain the concentration of each species present. If this operation is repeated at different values of pH, a speciation diagram is obtained from which it is possible to calculate formation constants using the formation reaction equilibrium as depicted in Equations 1.1 and 1.2. A graphical representation of the whole process is depicted in Figure 1.5.

$$p \cdot [MO_4]^{2-} + q \cdot H^+ \rightleftarrows [H_z M_p O_{4p-w}] + w \cdot H_2O; \quad w = \frac{q-z}{2} \tag{1.1}$$

$$r \cdot [XO_4]^{n-} + p \cdot [MO_4]^{2-} + q \cdot H^+ \rightleftarrows [H_z X_r M_p O_{4p+4r-w}] + w \cdot H_2O \qquad w = \frac{q-z}{2} \tag{1.2}$$

Some examples where this methodology was applied to phosphomolybdates can be found in the work of Pettersson et. al.[28] (see Figure 1.6) and later in the work of Cadot et. al.[39] For other kind of systems for which the metal nuclei is not active in NMR spectroscopy, ^{17}O-NMR has been used[40].

Among other techniques used to characterize polyoxometalates (see Figure 1.7) electronic and vibrational spectroscopy stand out. FT-IR[41–43] and Raman spectroscopy play a key role in the structural characterization of POMs. Raman spectroscopy is mainly used in aqueous solutions and IR spectroscopy can be used in aqueous and non-aqueous solutions. Due to the d^0 distinctive electronic structure of polyoxometalates, only a few absorption bands are expected in the UV-VIS[42] region of the spectra, related to charge transfer transitions. This technique is not able to provide structural information, but instead is used to determine the stability of polyoxometalates. To determine the structure of POMs some X-ray derived spectroscopies such as XAS[44] (X-ray absorption), SAXS[45,46] (small-angle

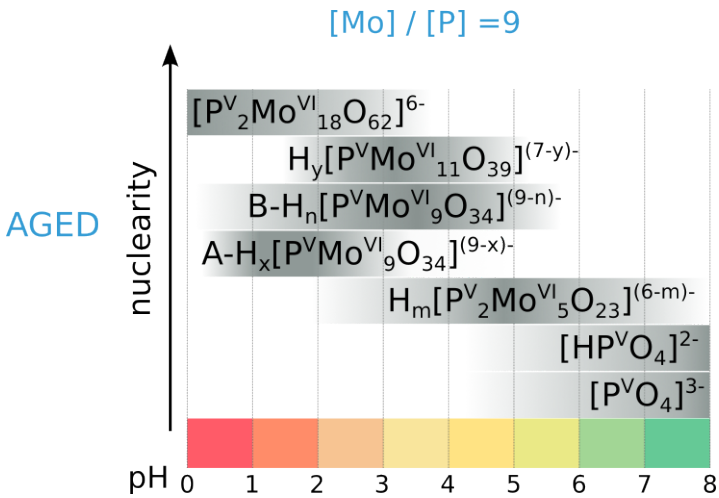

Figure 1.6: Figure adapted from reference [14]. Speciation of phosphomolybdates in aged (up to 1 month) aqueous solution with the concentration of 0.18 M Mo^{VI} ([Mo]/[P] = 9) based on works.167,168 The maximum intensity of grey color in each box with a single species corresponds to its maximum concentration in the chosen pH region. The grey boxes along the y-axis are positioned according to increasing nuclearity, but do not show the domination over other species at a certain pH range. The x value in $A - H_x[P^V Mo_9^{VI} O_{34}]^{(9-x)-}$ is 5–6; y in $H_y[P^V Mo_{11}^{VI} O_{39}]^{(7-y)-}$ is 0–2; n in $B - H_x[P^V Mo_9^{VI} O_{34}]^{(9-x)-}$ is 0–3; m in $H_m[P_2^V Mo_5^{VI} O_{23}]^{(6-m)-}$ is 0–2.

X-ray scattering), EXAFS[47,48] (extended X-ray absorption fine structure) or XANES[49] (X-ray absorption near edge structure) have been used. All these techniques provide structural information, but the SAXS is highlighted above the rest due to it's non destructive capabilities to determine the shape, the size or the reactivity of POMs. Cyclic voltammetry[50] is used to study the Red-Ox capabilities of polyoxometalates.

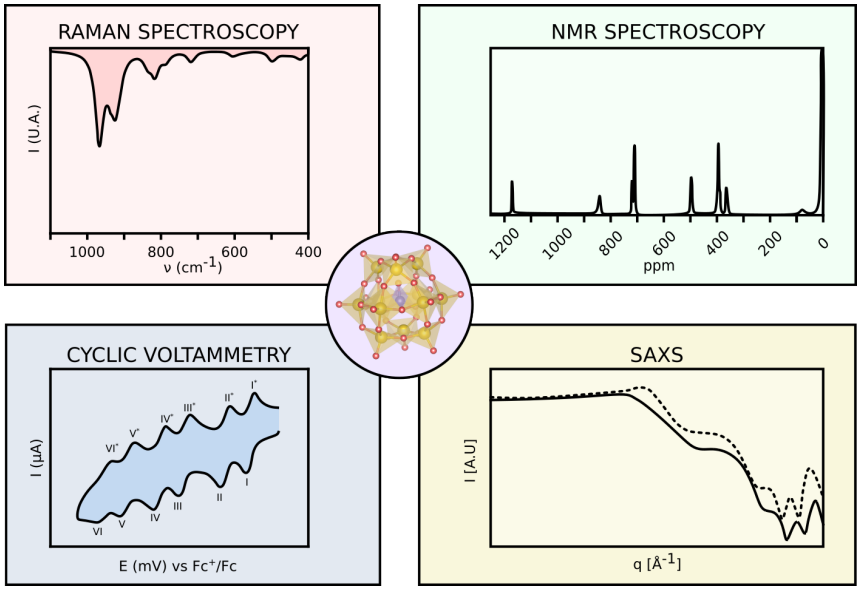

Figure 1.7: Exemplification of the most common experimental techniques used in the characterization of polyoxometalates: Raman and NMR spectroscopies, cyclic voltammetry and Small Angle X-ray scattering.

1.3 Computational studies on polyoxometalates

Although most of the pioneer advancements in the field depicted in Figure 1.2 correspond to experimental studies, computational approaches have gained importance since 1986[51]. Before that, the size of POMs and the amount of metals they contained made it impossible from a computational point of view. With the constant improvement of hardware and computational power, performing computational calculations on polyoxometalates started to become a reality. A timeline with some of the most important achievements in the field is depicted in 1.8.

At the beginning semi-empirical methods such as SCF CNDO-2 were

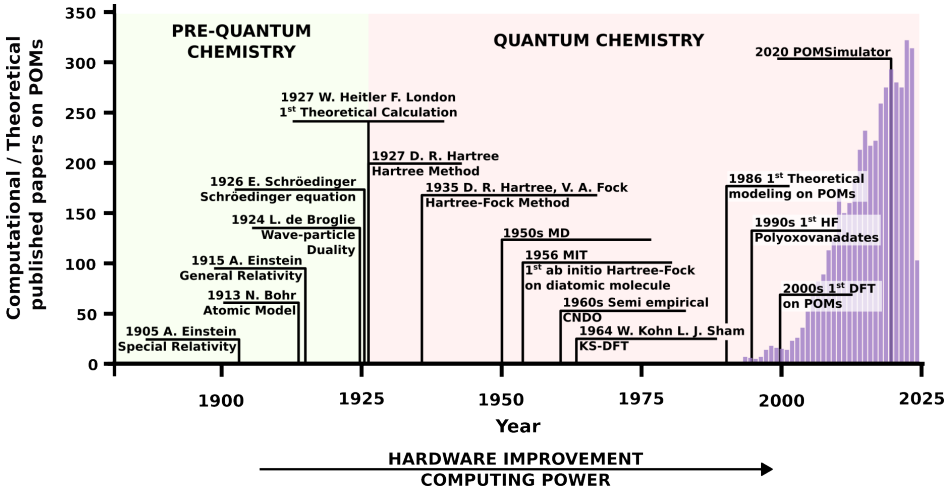

Figure 1.8: Chronological axis of the most important milestones in the history of computational chemistry and most important milestones related to POMs from the computational/theoretical point of view. The number of published papers related to polyoxometalates is shown as a line plot, with its y axis in the left side. The x axis corresponds to the time scale. (Query data for the line plot: [52]).

employed to study the electronic structure of polyoxometalates[53]. Later in the early 1990s, the first[54] *ab initio* Hartree-Fock (HF) calculations were performed increasing the knowledge acquired through semi-empirical methods. With the new wave-functions introduced in the HF method it was possible to further study the molecular electrostatic potential (MEP) and study the relative basicity of the oxygen atoms in POMs. But the biggest breakthrough in the field was the use of DFT[55] with POMs in the late 90s, and it is still the most used method nowadays. The introduction of different solvent models or considering relativistic effects improved the quality and accuracy of calculations and thus moved the field towards a proper modelling of polyoxometalates.

POMs computational studies have focused on different fields such as:

electronic structure[56–58], reactivity[59], MEPs[60], redox properties[61] and spectroscopy. IR, UV or NMR[40,62] have become important tools for the structural characterization of polyoxometalates as aforementioned in the previous section[63,64]. Experimental structural characterization often needs the help of computational tools to elucidate spectroscopic data.

In parallel, classical molecular dynamics (MD) have also been used to study POMs[65–67]. Using Newtonian equations of motion allowed the possibility to study the evolution of complex systems such as POM - solvent - counter cation through time. Since 1997 when the firsts POM FF were developed they have also been employed in the study of polyoxometalates[68,69]. In 2011 a new set of improved FF were developed which allowed to study the interaction of POMs with organic molecules[70]. Concerning the reaction mechanism of the self-assembly of polyoxometalates (POMs), several attempts have been made to elucidate this process.

Ab initio Molecular Dynamics (AIMD) was used in the study of nucleation mechanism of simple polyoxometalates into larger and more complex clusters[71]. Density functional theory (DFT) has been used to compute the bottom-up standard reaction mechanism from monomer for Lindqvist and keggin tungsten anions[71]. However, POMs are highly charged systems, and their stability is strongly dependent on pH. Including pH as a factor in all building blocks presents a significant challenge. Dr. Enric Petrus et al. successfully developed a method to simulate the aqueous speciation of iso-polyoxometalates[72], taking into account both pH and concentration. In *Chapter 2* it is briefly explained the functioning of the POMSimulator methodology. Figure 1.9 depicts a summary of some of the most employed QM/MD application in computational studies of polyoxometalates.

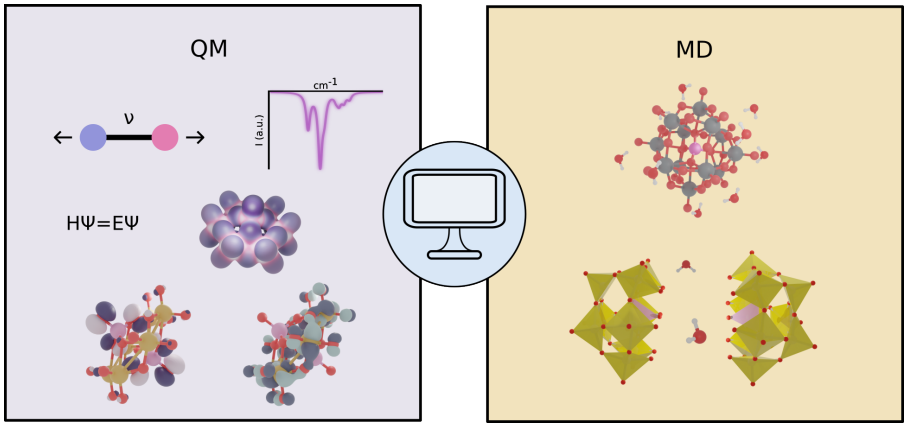

Figure 1.9: Graphical exemplification of some of the applications of QM and MD on polyoxometalates. In the DFT field some of the most important applications are: vibrational spectroscopy studies, electronic structure of electrostatic potential surface studies. In the MD field, aggregation processes or solvent interactions are some of the applications.

1.4 Objectives

The main target of this thesis is to further develop the POMSimulator framework to generalise the methodology, to improve the general performance of the code and to study the aqueous speciation of any kind of polyoxometalate, IPAs and HPAs. This main goal can be subdivided into smaller objectives, that have been worked in the different chapters:

- Adapt the POMSimulator algorithm for generating Chemical Reaction Networks (CRN) and adapt the solver for the system of equations for heteropolyoxometalates.
- Develop a universal scaling methodology for POMs to limit the dependency of the method on experimental results.
- Apply statistical techniques to improve the robustness of the method.

- Test and validate the new methodologies on the original studied systems.
- Apply the new approaches on HPA systems to obtain insights on their speciation.

Chapter 2 summarises the theoretical basis of POMSimulator, the different functionalities within its framework and the previous results.

Chapter 3 contains the method development of this thesis. It can be divided into three clearly defined sections:

1. Adaptation of the method for HPAs.
2. Implementation of a general scaling algorithm based on previous results including a Multi-Linear Regression (MLR) model and validation of the methodology.
3. Development of a new statistical workflow to classify speciation models and deal with large amounts of data and validation of this new methodology with the IPA systems.

Chapters 4 and 5 analyse the chemistry of the phosphomolybdate and arsenomolybdate respectively. These chapters have been structured similarly: defining the molecular set, describing the speciation results and the analysis of the CRN.

Chapter 2

Theory and method background: POMSimulator

2.1 Introduction

The complex aqueous speciation of polyoxometalates has been studied for many years, with significant advancements made over time. Research has demonstrated that POMs typically form through a series of well-defined steps, where smaller units progressively combine into more complex structures. Understanding the intermediate structures that emerge during self-assembly is essential for optimizing synthetic pathways and producing specific POM architectures. Efforts have been made to track how small metal-oxo units combine step by step into larger clusters. Using analytical techniques such as nuclear magnetic resonance (NMR), X-ray diffraction (XRD), and mass spectrometry, researchers can monitor these processes in real time, revealing critical details about how POMs grow and evolve.

NMR spectroscopy, in particular, has become fundamental in identifying key species that play crucial roles in solution[14]. Experimental and computational studies have significantly increased our understanding of polyoxometalates, their stability, and even their electronic structure. However, we are still far from fully understanding the behavior of POMs in solution. Several complex factors influence the self-assembly process, synthesis, and related speciation mechanisms. For instance, pH plays a crucial role, with acidic conditions often driving the assembly of larger clusters by promoting the condensation of smaller anionic units. Similarly, ionic strength and the concentration of metal precursors impact the size and shape of the resulting polyoxometalate clusters.

Five years ago, our group introduced a novel methodology called POMSimulator[72], which was able to simulate the aqueous speciation of molybdenum polyoxometalates. This new method relies on three fundamental pillars: quantum mechanics, graph theory and chemical equilibrium theory.

In this chapter, only the fundamentals of POMSimulator will be summarised as the full description of the method can be found in the doctoral thesis of Enric Petrus Perez and in the previous published articles[72–75].

The overall POMSimulator workflow can be described as:

1. Molecular Set: DFT calculations are carried out to obtain the energy and connectivity of the molecules to determine the bonds between the atoms.

2. CRN: Using Graph theory and stoichiometric relationships, the Chemical Reaction Network is set up.

3. Speciation models: Build determined solvable systems of equations.

4. Formation Constants: Solve speciation models and obtain formation constants.

5. Speciation: Generate speciation and speciation phase diagrams for the studied system.

2.1.1 Quantum mechanics: DFT

As aforementioned, POMSimulator relies on electronic structure methods to run.

Specifically, we make use of Density Functional Theory (DFT) to characterise the electronic structure of building blocks in the selected system and to optimize the corresponding geometries. DFT was first developed by Walter Kohn and Pierre Hohenberg in 1964 when they proposed the Hohenberg-Kohn theorem (HK)[76]. In this theorem it is demonstrated that the energy of the ground state of a given electronic system is a functional of the electron density, which depends on three spatial coordinates only. Later in 1965, Walter Kohn and Lu Jeu Sham proposed the Kohn-Sham DFT (KS-DFT)[77] method that solves the problem of electrons interacting in a static potential by considering non-interacting electrons in an effective potential. The KS equation is then solved iteratively, starting with an initial guess and solving until convergence, same as the Self-Consistent Field in the Hartree-Fock method. The true form of the functional proposed by Hohenberg-Kohn is yet unknown, for this reason approximate Exchange-Correlation functionals have to be proposed. Through the years, different functionals have been developed increasing the sofistication. In light of this in the year 2000, J. Perdew proposed the Jacob's Ladder[78] to classify the

functionals from less accurate (bottom part) to more accurate (top part). A schematic depiction of Jacob's ladder is given in Figure 2.1.

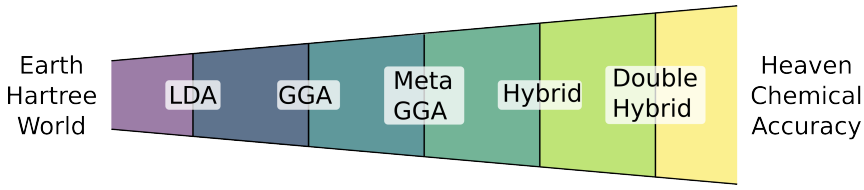

Figure 2.1: Representation of Jacob's Ladder.

In addition to the energies, the POMSimulator methodology relies on molecular connectivity. For a proper description of the molecular connectivity, Bader's *Quantum Theory of Atoms in Molecules*[79] is employed. The QTAIM analysis studies the topology of the electronic density of molecules and from this analysis it is possible to extract the critical points. The critical points are defined as the points where the first derivative is null, and thus they determine the positions of relative extremes in the electronic density (minima, maxima, saddle points). Critical points can be classified into Atom, Bond, Ring or Cage depending on the topology of the electronic structure.

In Figure 2.2, the QTAIM analysis performed on the $\{Mo_3O_9\}$ molecule is depicted. In this analysis, different critical points can be observed: bond critical points and ring critical points. Atom critical points remain hidden inside the atoms spheres. The contour lines represent the electronic density in the ring plane.

As it will be mentioned in the following section, it is crucial that all the molecules have proper connectivities as graph properties are sensitive to the graph topology.

The computational details related to the calculations carried out for

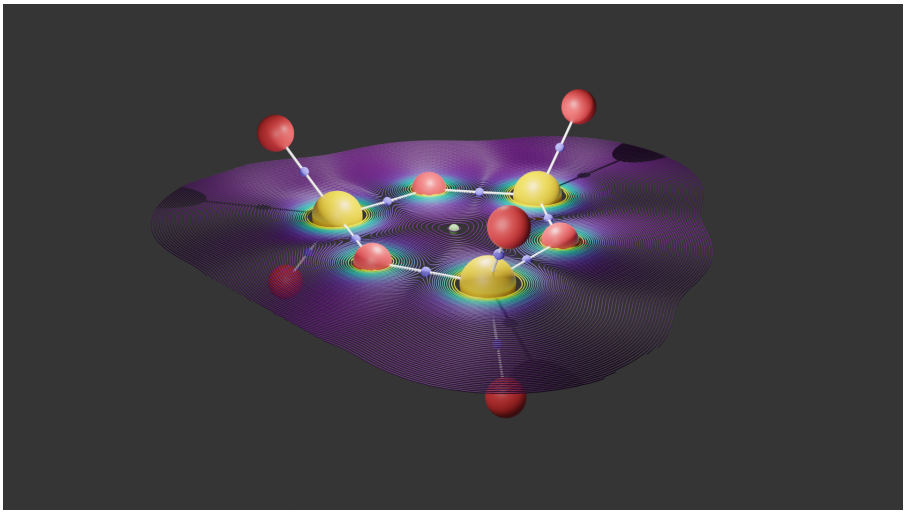

Figure 2.2: Bader's QTAIM analysis on the Mo_3O_9 molecule. Bond critical points are depicted as blue spheres. Molybdenum atoms are represented by yellow spheres, and oxygen atoms are represented by red spheres.

Table 2.1: QTAIM's classification of critical points of the electronic density.

CP types	
(3,-3)	Atom
(3,-1)	Bond
(3,+1)	Ring
(3,+3)	Cage

this thesis can be found on *Appendix A*.

2.1.2 Graph Theory

Once the essential data for all the metal-oxo clusters in the set has been gathered, the next step is to generate a reaction network that interconnects the multiple species. To address this task, we rely on the topological properties defined within Graph Theory and on predefined reaction types.

Graph Theory was first introduced by Leonhard Euler in 1736[80]. Graph Theory is by definition the study of graphs which are mathematical structures used to model relations between objects. Graphs can be defined as a set of nodes or vertices (N) and edges (E):

$$G = (N, E) \tag{2.1}$$

Graph Theory is used in a wide range of applications, from computer science to physics and chemistry. In this sense, graphs are a powerful representation of the molecular structure, as they can simplify complex structures into a set of nodes and edges, keeping important information as attributes, that can be stored in the nodes, the edges or even in the graph itself.

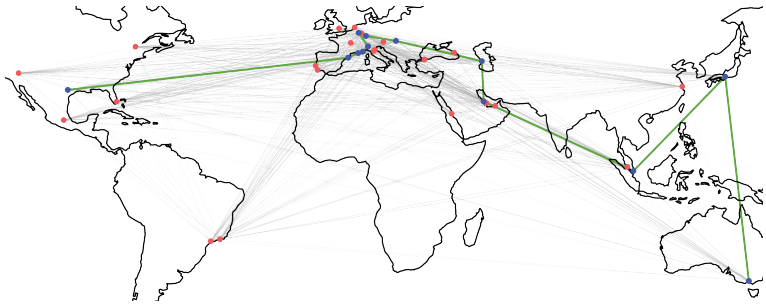

Figure 2.3: World map of the F1 circuits. Red and blue dots correspond to actual circuits, and the lines correspond to trips from one circuit to another.

An example of an application of graph theory to a real life problem can be found in the Formula 1 (F1) championship. For the past few years, the F1 has been trying to reduce their ecological fingerprint, reducing the amount of tyres used on race weekends or even using cleaner fuels. However, the traveling from one circuit to another is yet to be optimized to reduce the emissions of the F1. In this sense, finding the most optimal

calendar could be done by using graph theory. In Figure 2.3 the world map is depicted, with some of the actual circuits of the F1 championship. Using graph theory, it could be possible to find the least emissive calendar possible, storing the lengths of the trip in the graph edges, or even the carbon footprint.

In computational chemistry graphs can contain quantum properties such as energies or charges, and geometrical parameters such as distances or positions, making them very useful to analyze data. Any molecular structure can be represented as a chemical graph, where the atoms become nodes of the graph and the bonds become the edges. As an example, two different molecules have been drawn as graphs representations in Figure 2.4.

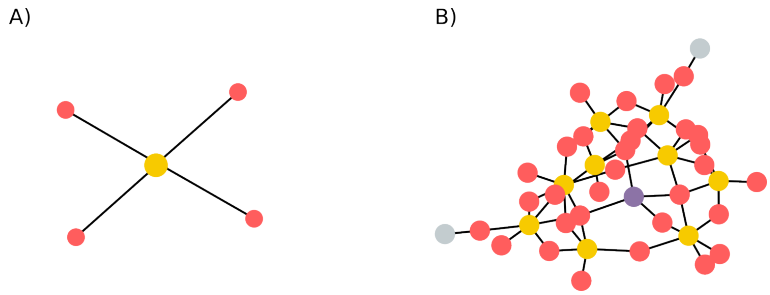

Figure 2.4: Graph representation of two different molecules, nodes are colored based on atomic number. Red color for oxygen, yellow for molybdenum, purple for phosphorus and grey for hydrogen. A) $[MoO_4]^{2-}$ molecular graph. It contains 5 nodes corresponding to 5 atoms, and 4 edges corresponding to 4 bonds. B) $[H_2PMo_9O_{31}]^{-}$ molecular graph.

In Figure 2.4, it's shown that the two graphs are quite different in complexity, so the graphical representation is affected by the size of the molecule and its inter-connectivity. If a molecule is linear-like, the molecular graph is less complex than the cluster-like structures, for instance a Keggin anion in Figure 2.4. On the other hand, graphs can be expressed as vectors or arrays, which can contain the graph attributes, the nodes and

their attributes and the bonds and their attributes Equation 2.2. By doing so, the representation of graphs gets simplified into a vector-like expression. These vectorial representations are very powerful as they can be applied to machine learning algorithms to train and predict properties of interest[81].

$$G = [\{G_{attrib}\}, \{nodes\{node_{attrib}\}\}, \{edges\{edge_{attrib}\}\}] \tag{2.2}$$

But the most important application of molecular graphs inside the POMSimulator framework is the ability to automatically build complex reaction networks (CRN). Chemical Reaction Network theory is an area of mathematics focused on modelling chemical systems. CRNs have proven to be an ideal tool when dealing with complex reaction networks systems[82–86].

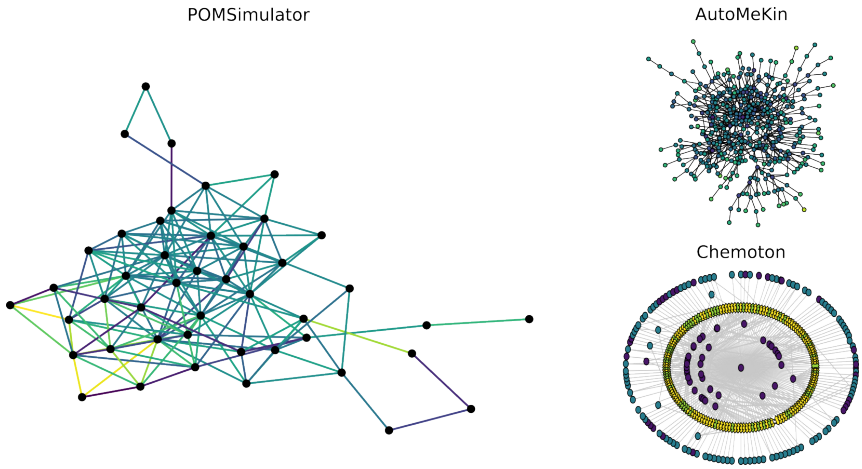

Figure 2.5: Examples of automated programs for generating CRNs. On the left side, an example of the complex CRN of the W system generated with POMSimulator. It includes 116 reactions and 45 polyoxotungstates. On the right side, AutoMeKin[87] (top) and Chemoton[88,89] (bottom).

In Figure 2.5, 3 different Chemical Reaction Networks generated with different methodologies are depicted. Through molecular graphs and their

properties, POMSimulator is able to generate complex reaction networks for the POMs in study in an automatic way. In our case, we represent the CRN as graph of molecular graphs, in which the nodes are molecules and the edges are the reactions. To determine the reactions present in our system, we will make use of subgraph isomorphism property and stoichiometry.

The subgraph isomorphism will determine whether a graph G_2 is formed from a subset of the vertices and edges of another graph G_1. In this sense a subgraph isomorphism matrix is calculated beforehand. As it is not important if G_1 fulfills the subgraph isomorphism of G_2 or G_2 of G_1, this square matrix is transformed into a triangular matrix, only checking each pair of graphs once. In this matrix *1s* represent pairs of molecules that accomplish this property, and *0s* represents pairs of molecules that don't.

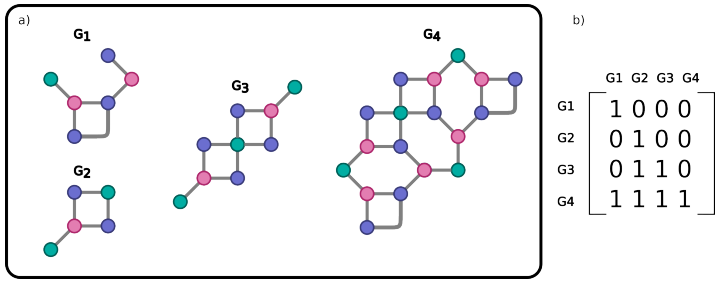

Figure 2.6: a) Representation of four distinct graphs $G_1 - G_4$ showcasing different connectivities.G_1 only fulfills the subgraph isomorphism with G_4, whilst G_2 fulfills it with G_3 and G_4 at the same time. b) Subgraph isomorpgism trigonal matrix. Each pair of graphs is checked only one.

For example, in Figure 2.6, G_1 and G_4 would be related by the subgraph isomorphism property, as G_1 is a subset of nodes and edges of G_4. On the other hand, G_2 would fulfill this property with both G_3 and G_4. As this property relies on the topology of the molecular graphs, it is important to define the molecular connectivity properly. In our case, we employ Bader's

QTAIM[79] analysis to determine the bond critical points in the electronic density of the molecules to establish the connectivity between atoms.

The second condition to detect a chemical reaction in POMSimulator is the stoichiometry. We pre-defined different reaction types related to polyoxometalates chemistry such as acid-base reactions, condensations, additions among others. After the subgraph isomorphism is checked, the molecules that fulfill that property will be checked for stoichiometric balance. This balance is expressed as a list of each of the present atoms unique balance. Isopolyoxometalates are formed mainly by a metal (M) and oxygen (O), but can also contain protons (H). From the point of view of POMSimulator, the stoichiometry of poms is expressed as [M,O,H]. Thus, the atom balance is also represented as [ΔM,ΔO,ΔH]. The resultant stoichiometry is then categorized in one of the pre-defined types according to Table 2.2. As a result of both, graph topology and reaction type, the chemical reactions that relate our chemical graphs is defined. This set of chemical reactions will define the Chemical Reaction Network of the system, and the complexity of it. With this CRN we will now be able to generate the system of equations to generate speciation models.

Table 2.2: Reaction types definitions. First and second columns represent the reaction type followed by a generic reaction. The last column represents the stoichiometric balance of the reaction. This balance is calculated by a diference between the product and the reagents.)

Reac. Type	Reaction	Stoichiometric Balance
Acid-Base	$A + H_3O^+ \rightarrow AH + H_2O$	[0,0,1]
Condensation	$A + B \rightarrow C + n \cdot H_2O$	n· [0,-1,-2]
Addition	$A + B \rightarrow C$	[0,0,0]
Hydration	$A + n \cdot H_2O \rightarrow B$	n· [0,1,2]

2.1.3 Chemical Equilibrium

Once the CRN is constructed, a system of equations is set up to mathematically solve the speciation of the system. For polyoxometalates or systems with nucleation reactions like polymerizations, the number of chemical reactions is higher than the number of species, which results in an over-determined system. To solve this complexity we generate what we call speciation models (SM). We define a speciation model as a subset of reactions that added to a mass balance equation generate a solvable determined system of equations. In Figure 2.7, an example of the construction of a SM is depicted. In this example, the selected system consisted of a total of 10 chemical reactions and a mass balance, and a total of 6 species. As previously mentioned, the system is over-determined, so SMs are needed to solve this system.

To achieve this, a subset of equations is selected to ensure that the number of equations matches the number of species in the system.

The problem here is the combinatorial explosion associated with the number of possible speciation models. From a simple set of 11 equations, the possible combinations of 6 elements is 462 ($C = \binom{n}{k} = \frac{n!}{k!(n-k)}!$). If we added more equations, the number of possibilities would grow factorially, making it a severe problem.
To solve it, two hypotheses were proposed:

1. All speciation models had to include the acid-base reactions, as the polyoxometalates chemistry is strongly pH dependent.

2. Each speciation model had to include a nucleation reaction to form each nuclearity (defined as all protonation states of the same species).

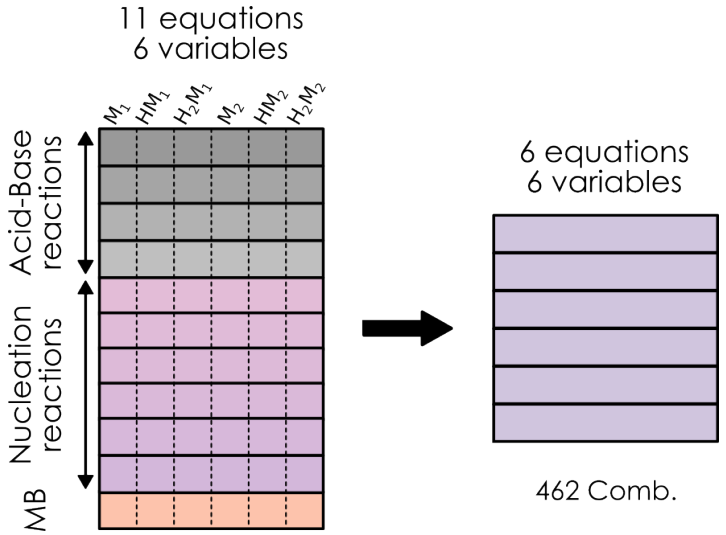

Figure 2.7: Graphical representation of the construction of 1 speciation model.

In Figure 2.8 an schematic representation of the construction of SM applying the two hypotheses is depicted.

In Table 2.3 we compare the previous example with the phosphomolybdate (PMo) system. The number of speciation models calculated as the combinatorial is around 10^{31} for the PMo, although after applying the two conditions this number decreases to around 300 million models. This is important, as computationally speaking, it is impossible to solve 10^{31} speciation models in a reasonable amount of time. To solve the system of equations we need to set a grid of pH values and solve the system of equations at each point of pH. The solution of the system of equations is the concentration of all the molecules across the pH grid. With these concentrations POMSimulator is able to calculate the formation constant (Kf) for each molecule, as it is detailed in the following section. It's worth noting

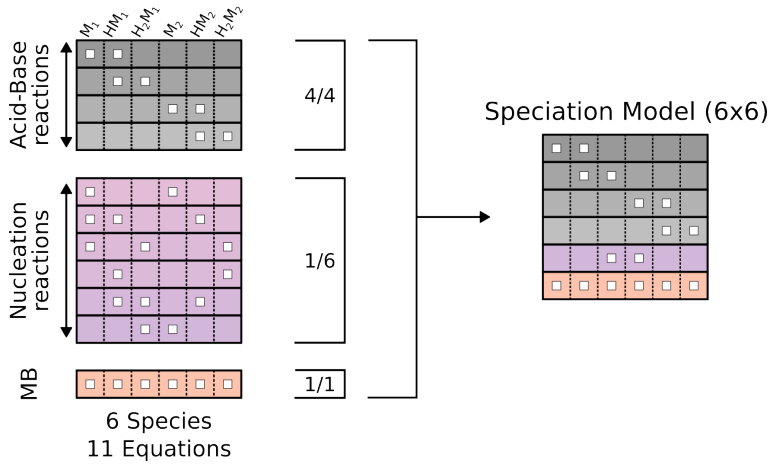

Figure 2.8: General construction of speciation models applying chemical-based hypotheses. Species involved in each reaction are highlighted with an empty square.

Table 2.3: Comparison between the generation of speciation models, considering the mathematical combination or applying chemical assumptions.

	Total Spec. Models	
	6 spec. 11 reac.	49 spec. 109 reac.
Mathematical Comb.	462	$2.85 \cdot 10^{31}$
Chem. Assumptions	6	$\approx 3 \cdot 10^{8}$

that each speciation model results in different formation constants for each specie.

2.2 Chemical speciation

The comparison between the computed formation constants and the experimental formation constants shows that the DFT-calculated values are overestimated and must be scaled to match experimental data accurately. By applying scaling and making linear correlations, the root mean square

error (RMSE) can be calculated (Figure 2.9). The speciation models are then ranked according to their respective RMSE values, and the model with the lowest error is selected as the best model. In the initial release of POMSimulator, this method was used to choose the speciation model with the lowest error. Once the best speciation model is selected, formation constants, its speciation diagram, phase diagram.

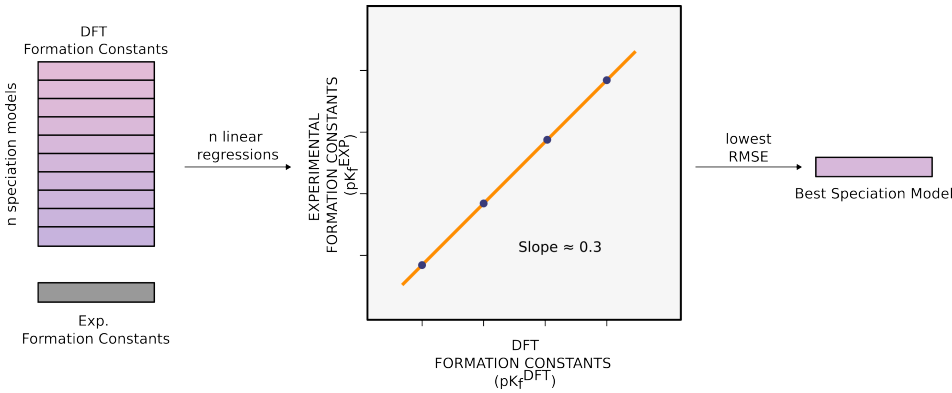

Figure 2.9: Example of Linear scaling and best model selection.

Once we have selected the best speciation model, its constants are scaled and their speciation diagram calculated. To do so, a new equations system must be set up which consist of a formation reaction for each species in the molecular set except for a reference compound, typically a monomeric species without any proton, with the following generic formula MO_4 or MO_3 in some cases. Formation reactions will follow the pattern of Equation 2.3, and according to this reaction we can establish its equilibrium constants Equation 2.4), in this case formation constant.

$$p \cdot [MO_4]^{2-} + q \cdot H^+ \rightleftarrows [H_z M_p O_{4p-w}] + w \cdot H_2O; \quad w = q - \frac{z}{2} \tag{2.3}$$

$$K_{eq} = K_f = \frac{[H_z M_p O_{4p-w}] \cdot [H_2O]^w}{[MO_4^{2-}]^p \cdot [H^+]^q} \tag{2.4}$$

Again, we will have a set of *N-1* equations being N the number of species, to which a mass balance equation has to be added. This way, for each speciation model, after solving a new equations' system, concentrations for all the species in the molecular set are obtained, which will later generate speciation diagrams and phase diagrams (see Figure 2.10).

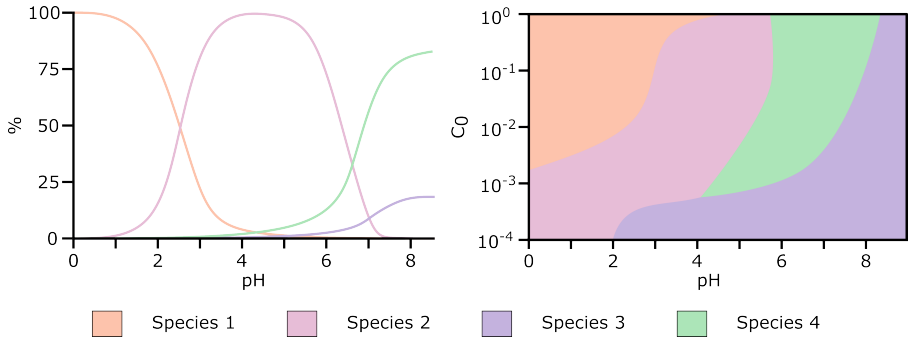

Figure 2.10: Example of speciation (left) and phase (right) diagrams. In speciation diagrams we see how species behave at a given initial concentration throughout different pH values. In phase diagrams we see how the initial concentration and pH affect the distribution of species. Only the predominant species at each pH, C pair are plotted.

Speciation diagrams represent the variation of the concentration the species across a pH range. This kind of diagrams are often used to describe the acid-base equilibria of different acids and bases, showing the distribution of the different protonation states across the pH. For the polyoxometalates, speciation diagrams describe at each pH point which are the concentration of the species in solution. On the other hand, speciation phase diagrams can be described as a concatenation of different speciation diagrams. While the speciation diagrams show the evolution of concentration at different pH values, the phase diagrams show the predominant species at a given set of

conditions (initial concentration, pH, ionic strength, temperature...). The way speciation phase diagrams are generated within the POMSimulator framework is by collecting speciation diagrams generated at different initial concentrations of metal.

2.3 Application to IPAs

This methodology was successfully applied to 5 different isopolyoxometalate systems: Mo and W first[73] and V, Nb and Ta second[74]. Initially, POMSimulator was applied to the Mo_8 system[72], and since then, the complexity of the studied systems grew considerably. In the Mo system, it could be observed how a high concentration of metal and low pH gave place to the largest cluster $\{Mo_{36}\}$. Either lowering the initial concentration or increasing the pH resulted in the decomposition of the $\{Mo_{36}\}$ into smaller clusters such as $\{Mo_8\}$ or $\{Mo_7\}$. At even lower concentrations or higher pH, only monomeric species at different protonations states were present. In the W system, the $\{W_{10}\}$ was present at any initial concentration in a pH range between 1 and 4.5. At higher pH values, the $\{W_{12}\}$ was also present at any initial concentration. Regarding the W system, our group studied the kinetics of the transformation of the $\{W_{12}\}$ into $\{W_{10}\}$[75]. For both the Mo and the W, the speciation took place in acidic media and in agreement with experimental results (Figure 2.11).

Shifting the focus to the group 5 chemistry, the speciation of niobium and tantalum took place in more basic conditions (Figure 2.12). In between the acidic conditions for Mo and W, and Nb and Ta, the speciation for V took place in a neutral range of pH, showing that POMSimulator can describe both basic and amphoteric chemistry in alignment with the

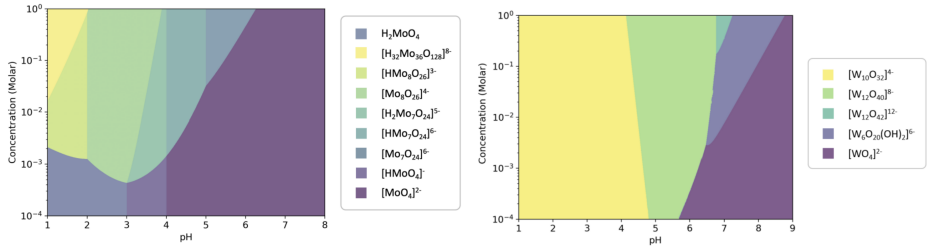

Figure 2.11: (Left) Speciation phase diagram for molybdenum system. (Right) Speciation phase diagram for W system. Image adapted from [73].

experimental results.

It is valuable to point out some more differences between various studied systems. For the V speciation, it is possible to see the formation of the metavanadates ($\{V_4\}$ and $\{V_5\}$) at neutral pH values in opposition with the previous systems in which no nucleation was appreciated at that pH value. At lower pH, the formation of $\{V_{10}\}$ in two protonation states can be appreciated. For the Nb and Ta systems the speciation was pretty similar in both cases.

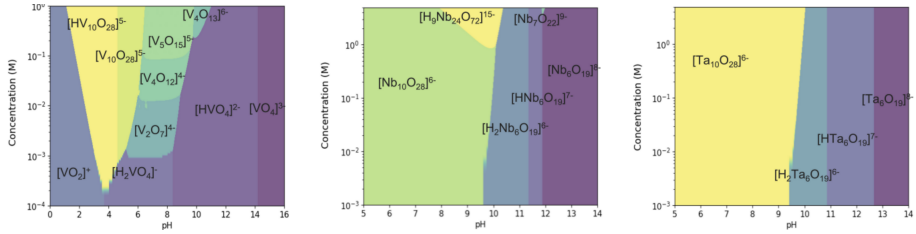

Figure 2.12: (Left) Speciation phase diagram for the V system. (Center) Speciation phase diagram for Nb system. (Right) Speciation phase diagram for Ta system. Image adapted from [74].

The $\{M_{10}\}$ was formed in pH values lower than 9, whilst the Lindqvist structure was formed from pH higher than 9. In the Nb system, there is a difference which is the $\{Nb_{24}\}$ structure which is form in very high

concentrations and slightly basic pH. In all cases it is possible to see how pH has proven to play a key role in the speciation.

These phase diagrams emphasize that not only the pH, but also the initial concentration of metals, play a crucial role in the speciation. For instance, lower initial concentrations can lead in some cases (Mo and V) to the absence of clusterization, or in the case of Nb, only in high concentrations it is possible to form the $\{Nb_{24}\}$.

The initial open release of POMSimulator was made publicly available on GitHub[90] following the FAIR data principles. With that release it is possible to reproduce the results on Mo, W, V, Nb and Ta[72–74].

Chapter 3

POMSimulator 2.0: Pushing the limits

After studying 5 different IPAs systems with POMSimulator, we turned our gaze to HPAs aiming to predict the formation mechanism of the Keggin anion. We initially selected the phosphomolybdate (PMo) system since it was the one that had the most experimental data available and thus would be the easiest to compare our results to. As mentioned in the previous chapter, the complexity of the systems under study has been constantly increasing which translates to a factorial growth in the number of speciation models. This continuous growth in complexity demanded a deeper and more thorough analysis. In this Chapter we propose two new methodologies to increase the predictive power of POMSimulator as well as increasing the robustness through a statistical analysis of the speciation models.

3.1 Combinatorial explosion of speciation models

The inclusion of a heteroatom in the system requires the modification of the way we generate the CRN and the way we build the speciation models. As introduced in the previous chapter, we generate the reaction networks through graphs comparisons and stoichiometric difference according to a pre-defined set of reaction types. Introducing a new atom type entails a modification of these reaction types. Also, we need to include a new mass balance equation to account for the newly included heteroatom, and in the same way we did with IPAs, we need to include in each speciation model a reaction that forms each of the nuclearities present in the set, except for the reference compounds (generally monomeric species of both the metal and the heteroatom).

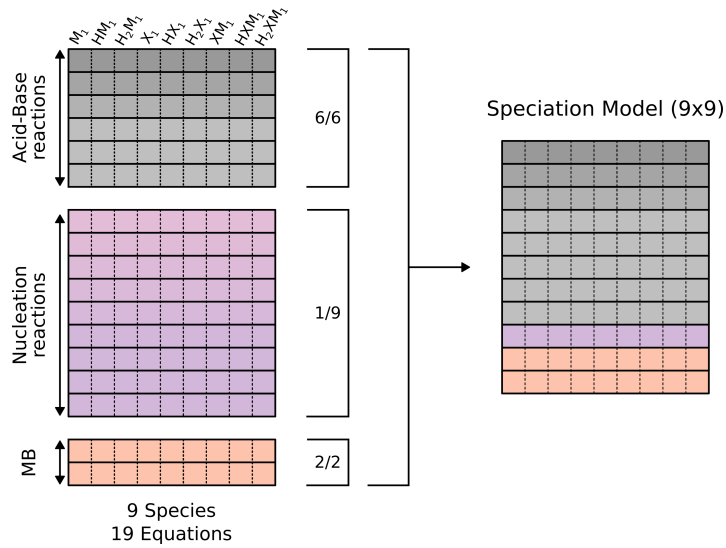

Figure 3.1: Schematic depiction of the general formation of speciation models for HPA systems. Two mass balance equations must be considered, one for the metal and one for the heteroatom.

Figure 3.1 depicts a general scheme of speciation models for the most straightforward HPA system example possible. This example comprises the monomeric species of the metal and the heteroatom (M_1 and X_1) in different protonation states and a species containing one metal and one heteroatom. In a first look, it is appreciable than in comparison with the simplest IPAs system, HPAs create a more complex system of equations, that gets even more complex when scaled to real systems. For instance, the PMo systems consisted of 49 species and 109 reactions which was transformed into approximately 300 million speciation models.

Table 3.1: Comparative between different POM systems. The secon column represent the number of species considered in the molecular set. The third column accounts for the number of reactions considered for each system. The last column corresponds to the number of speciation models calculated by POMSimulator.

Metal	N. spec.	N. reacs.	N. SM
W	51	67	50k
Nb	39	66	500k
V	42	75	1M
PMo	**49**	**109**	**300M**

In Table 3.1 a comparison between the PMo system and the previous studied systems is depicted. The biggest system so far was the vanadium one with around 1 million models which compared to the PMo (around 300M) differs by a factor of 10^2. This increase is even higher if we compare with the smallest system which defers by a factor of $6 \cdot 10^3$.
Hardware and method limitations forced the first big change in POMSimulator. We couldn't solve the totality of SMs generated in the PMo system (around 300 millions). Given the impossibility to solve the totality of speciation models, we aimed to explore if sampling a fraction of the total SMs would be applicable. This new concept prompted some questions:

1. Would a small % of the total models be statistically representative?
2. Could we miss the best model if we don't calculate all the speciation models?
3. Would considering only the best model be enough to describe a complex system with millions of speciation models?

All these questions implied some important changes into the POMSimulator methodology to be carried out in order to be answered.

The first modification to solve SMs in large systems with POMSimulator was applying a random sampling to the set of SMs and select a solvable amount of models. To deal with the PMo system we applied the random sampling to the set of SMs and selected "only" 1% (3M out of 300M). This sample was selected randomly and consisted of 3 million speciation models, three times the largest system studied in POMSimulator. The sampling had to be randomly performed, due to the way speciation models are generated. Adjacent SMs differ very slightly, possibly only a single reaction is different, and for this reason, if we only selected the first 1% we wouldn't be considering the diversity across SMs.

As a result of the first change, the concept of best speciation models had to be revised. We needed to find a different way to treat the results obtained from POMSimulator. So far, we were only considering the best speciation model (based on the RMSE scoring from the linear regresion) for each of the studied systems. After applying a random sampling, the best model could be lost in the uncalculated speciation models and for this reason we realised that we should consider new solutions.

Originally, the best model was used for the prediction of the speciation diagrams and the speciation phase diagram, and to do so, the calculated

formation constants had to be properly scaled. As it has been previously mentioned, the linear scaling is performed by comparing the raw DFT formation constants calculated with POMSimulator and the available experimental formation constants. Therefore, the selection of speciation models is entirely linked to the formation constants.
After applying POMSimulator to five different IPA systems[72–74], it was seen that the slope parameter of the linear regression for all the systems was around 0.3 but the intercept was variable across the different systems. This behavior led us to think that there could exist a general or universal scaling for formation constants. In the following Section 3.3, a methodology based on Multi-Linear Regression model to predict the scaling parameter is detailed . We then developed a new methodology based on classifying speciation diagrams through clustering techniques and standard statistics which allowed us to properly predict the experimental[14,28,39] speciation of the PMo system. This methodology is thoroughly explained in Section 3.4.

3.2 Linear scaling of formation constants

The calculated formation constants of polyoxometalates are overestimated if compared to experimental values. For this reason the initial approximation to solve this problem was to perform a linear regression between the calculated formation constants and the experimental formation constants. Using the parameters of this regression (slope and intercept), the constants were scaled to their proper value. Originally, for each speciation model, a linear regression was performed and models were sorted according to their respective RMSE (root mean squared error). The model with lowest RMSE scoring was selected as *best model*. An interesting pattern

was found after studying five different IPAs. The slope of the regression for all systems and speciation models was around 0.3, hinting at a possible universal scaling factor for polyoxometalates. On the other hand, the intercept was variable across different metal systems.
In comparison to the original method, scaling all the models to a unique value, would considerably reduce computation cost due to not having to perform thousands of linear regressions. Moreover, the methodology would be set free from experimental dependence, which is one of the bottlenecks of the POMSimulator methodology. The dependence on experimental results restricts POMSimulator to study only systems with available formation constants (of which there are not that many), when there are many interesting systems that would remain unstudied.

With the idea of solving a random sample of speciation models, selecting a best models doesn't make sense anymore. For this reason, calculating thousands of linear regressions also doesn't either make sense as it is time consuming. With all of this, coming up with a new way of dealing with the scaling of the calculated formation constants. The idea of having a unique scaling methodology for any system gained importance as it was observed that all studied systems had a similar scaling function ($K_{f_{exp}} \approx 0.3 \cdot K_{f_{DFT}} + b$).

To accomplish a unique scaling factor for all the POMs we gathered all the calculated constants for the five IPA systems, and the ones calculated for the PMo system. In order to make the new methodology more robust and independent on the amount of speciation models of the studied systems, the calculated formation constants of each were compressed into the median value across all the speciation models. This way there was only one single value for each species and for every system.

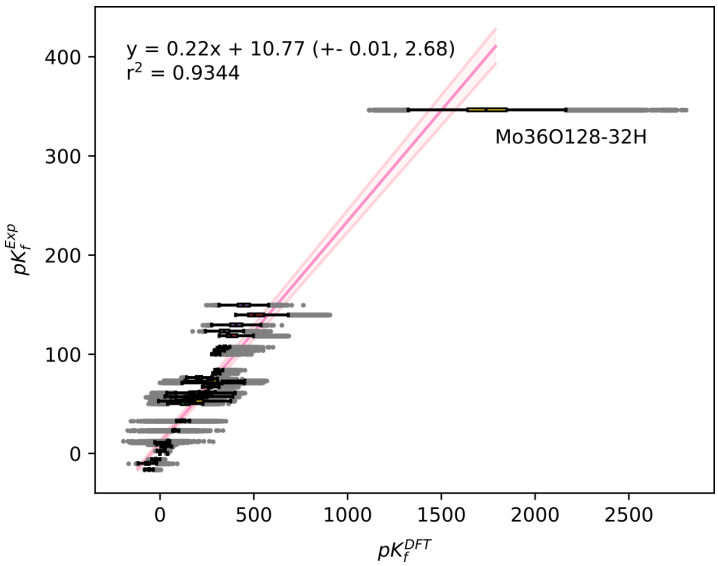

Figure 3.2: Linear regression for all species studied in POMSimulator that have available experimental formation constants. The DFT value for the formation constants of the different considered species is represented as box and whiskers plot, and the linear regression is performed with the median value of the DFT constant.

We used the aforementioned median value of the K_f to calculate a single linear regression with all systems at the same time. For a better visualization and understanding of the result, we plotted the calculated formation constant as a horizontal box plot centered in the y axis in their respective experimental formation constant (Figure 3.2).

During this Chapter, box and whiskers plots (boxplot) are widely employed as they are the perfect way to represent our data (formation constants, concentrations ...). Boxplots are used in descriptive statistics to graphically demonstrate the spread and skewness of numerical data through their quartiles. Boxplots usually contain lines called whiskers that extend the boxplot to indicate the variability outside the upper and lower quartiles.

The numerical data beyond the whiskers is called Outliers which are usually plotted as dots. Figure 3.3 depicts an example of a box and whiskers plot. In this figure, the general structure of a box and whiskers plot is described. The center line represent the median value, and the limits of the box represent the first (25%) and third (75%) quartile. The distance between Q_1 and Q_3 is called Inter Quartile Range (IQR). In the POMSimulator framework the whiskers are calculated as 1.5 times the IQR. Any data point outside the whiskers is considered an outlier.

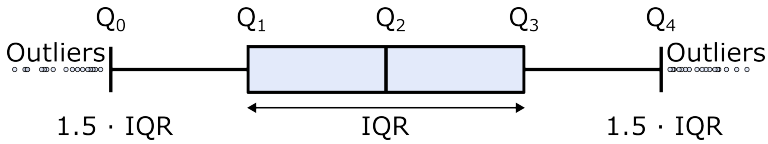

Figure 3.3: Scheme of the structure of a box and whiskers plot, with the corresponding statistical parameters.

The center pink line represents the linear regression, while the outer lighter pink lines represent the standard deviation (σ) as an error band. The slope value for this regression is far (0.22) from the observed $\approx$0.3 value in all the systems. For this reason we plotted each system separately, to find any outlier across the studied species. In Figure 3.4, a depiction of the regression of each system is presented. From this figure, we can appreciate that the IPA-Mo system presented a different behaviour than the other systems.

At this point, we wanted to determine if the molybdenum system was significantly different on its own, or if there was any species that was an outlier. To do so, we gathered all the species but one, and performed the linear scaling again. We repeated this operation until we had tried all the combinations of leave-one-out. For every combination we calculated the parameters (slope and intercept) from the linear regression and the r^2

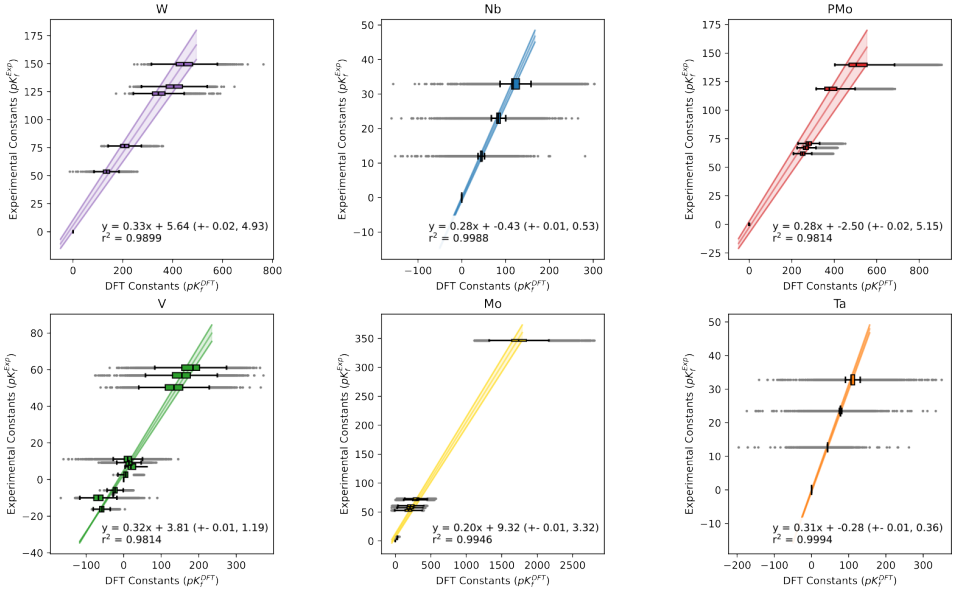

Figure 3.4: Linear regressions for all the species with experimental constants, separated in different panels according to the system they belong to. Again, the linear regressions are carried out using the median values of the DFT constant. For each metal system their respective scaling equation and r^2 scoring is given.

scoring. This way, for each species we leave out of the linear regression we obtain a set of slope and intercept parameters.

In Figure 3.5 there is a depiction of the results from all the combinations of leave-one-out. On the left side of the figure an scatter plot using the value of the slope (x axis) and the value of the intercept (y axis) is shown. From all the combinations of leave-one-out, dropping the $\{Mo_{36}\}$ from the regression shows the most distinct behaviour. In other words, all the combinations that contain the $\{Mo_{36}\}$ species show the same trend. The results obtained in Figure 3.4 demonstrate that except for the IPA-Mo system, all the calculated slope parameters were around 0.3. For this

reason, the results of the leave-one-out point in the same direction: only after removing the $\{Mo_{36}\}$ species, we obtain similar results for the general regression of all systems.

On the right side of the figure, we plot the value of the slope versus the r^2 scoring. In this case the same pattern is observed. While removing any specie from the set doesn't considerably change the overall regression, discarding the $\{Mo_{36}\}$ results in a change of behaviour.

From these results, we can conclude that the experimental formation constant for $\{Mo_{36}\}$ species is an outlier in the IPA-Mo system and for this reason we removed it from the set.

It is worth mentioning that the DFT calculations for this species were troublesome implying that the structure of this molecule might not be properly described[91].

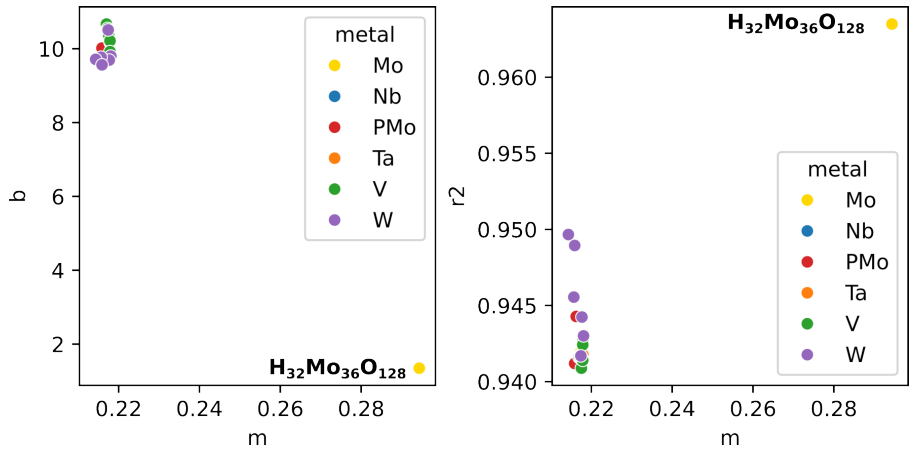

Figure 3.5: Application of the leave-one-out methodology to the linear regression from all the systems. Left panel represents the distribution of the slope parameter against the intercept for all the combinations of leave-one-out. The right panel represents the distribution of the slope value with the r^2 scoring value. Points are colored according to the system they belong.

After removing the outlier from the set, we can perform the linear re-

gression once again. The results from the new regression are depicted in Figure 3.6. The slope of the linear correlation now follows the same line as the individuals regressions.

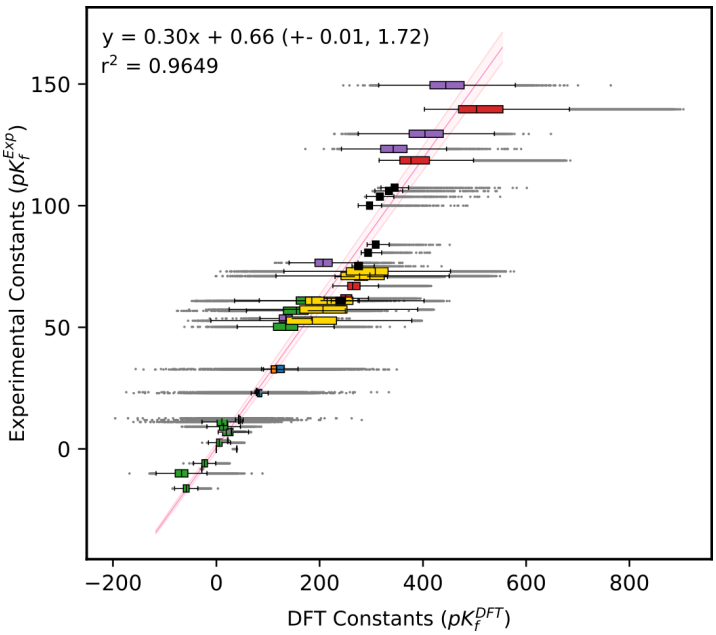

Figure 3.6: Linear regression with boxplots for all the species but $\{Mo_{36}\}$. Equation parameters differ from the ones obtained in Figure 3.2.

With these slope and intercept parameters, we scaled the formation constants calculated for a new HPA system, the arseno-molybdate (AsMo). The AsMo system might not be one as famous as other HPAs like the phosphomolybdate, the silicotungstate or the phosphotungstate, but it has been deeply studied in the past years[14,92–96].

Speciation diagrams at different concentration ratios for this system were calculated with disappointing results. To generate those speciation diagrams we employed the methodology described in Section 3.4 of this chap-

ter. A general trend was observed, nucleation reactions (only monomers acid-base chemistry) were not occurring in contrast with experimental results (Figure 3.7).
To find the error, we compared the average scaling parameters of the arsenomolybdate system with the universal scaling parameters that we used. The difference in the slope was minimal, but on the other hand, the difference in the intercept was considerable (around 8 logarithmic units). This huge discrepancy was playing a key role in the miscalculation of the speciation diagrams. While the slope parameter affects all constants in the same way, the intercept has bigger effects in smaller constants than in larger ones. For this reason, a further and deeper study of the calculation of the intercept had to be performed. The results of the speciation of the AsMo system are described in *Chapter 5*.

3.3 Universal scaling of formation constants

The advantages of having a unique scaling for formation constants of all polyoxometalates lie on the independence of experimental results, the universality of the POMSimulator methodology and the scalability of the method with the inclusion of new data.
As aforementioned, there is a problem in the calculation of the intercept parameter of the universal scaling linear regression. While the slope remains constant in a value close to 0.3, the intercept is quite sensitive to the studied system (3.4), ranging from 9.32 (Mo system) to -2.5 (PMo system). For this reason, we needed to develop an algorithm based on the calculated formation constants, which could predict the intercept parameter.

At first stage, in order to check the universality across DFT functionals

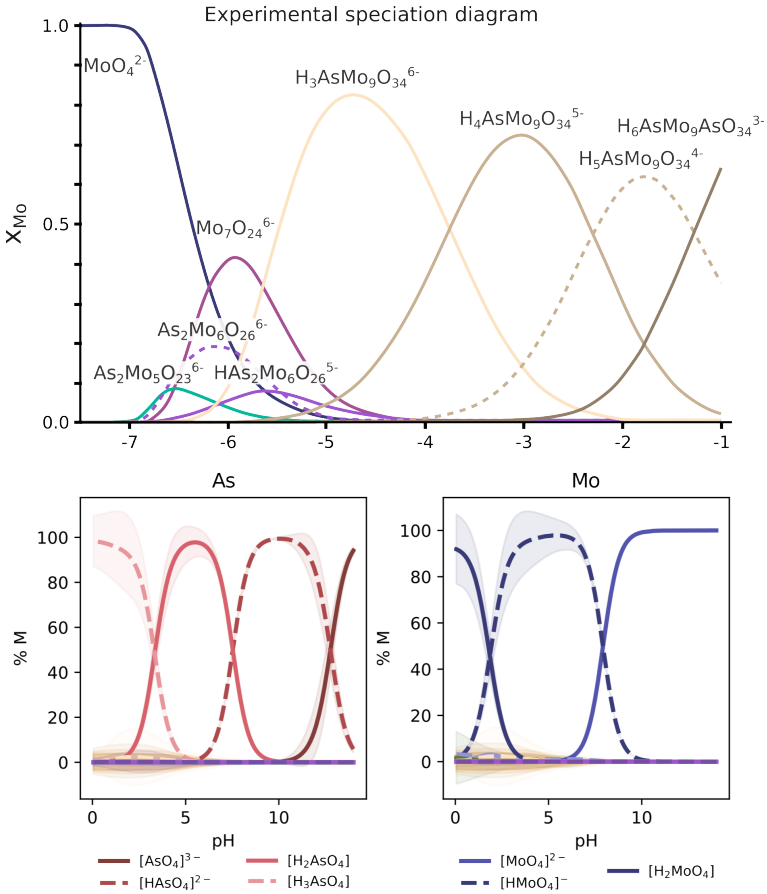

Figure 3.7: Comparison of the experimental[29] results and the results obtained in the speciation of the AsMo system. Parameters used to compute the speciation: temperature = 298.15K, C_{Mo} = 0.09M and C_{As} = 0.01M.

we reproduced the linear regression with all compounds using four different DFT functionals: PBE and BP86 (GGA), B3LYP (hybrid) and M06L (metaGGA). In all cases, the results were very similar in terms of slope and of intercept (Figure 3.8). For this reason, only PBE has been employed to

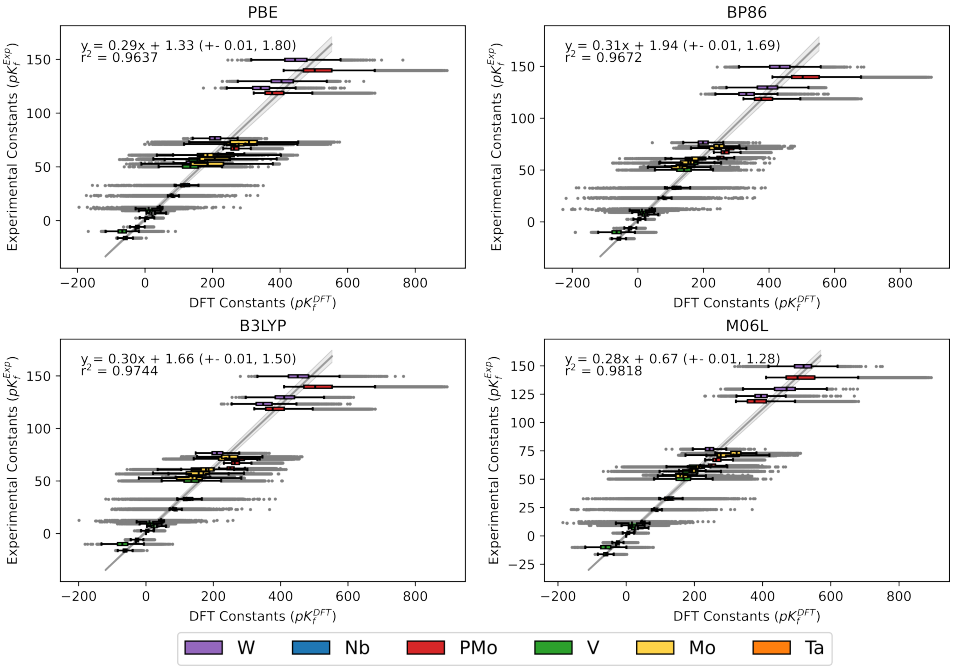

Figure 3.8: Linear regression for the universal scaling algorithm using different functionals.

develop the new scaling algorithm.

When examining the differences among the studied systems, it became evident that the intercept parameter needed to be system-specific. We exclusively used data derived from the POMSimulator methodology for predicting these scaling parameters. The calculated formation constants were an ideal input for the universal scaling protocol as they comprise all the complexity of the system. For each system, using POMSimulator we obtained a set of speciation models and for each of them a formation constant was obtained for every species in the molecular set. Considering the set of speciation models allowed us to calculate statistical descriptors such

as the mean ($\bar{x}$), standard deviation (σ), Q_1 and Q_3 quartiles, range, and minimum (Min.) and maximum (Max.) values for the formation constant of each species.

- $\bar{x}$: Is a numeric quantity representing the center of a collection of numbers. It is intermediate to the extreme value.
- σ: Is a measure of the amount of variation of the values of a variable about its mean value.
- Q_1: The first or lower quartile is the value under which 25% of data points are found when they are arranged in increasing order.
- Q_3: The third or upper quartile is the value under which 75% of data points are found when arranged in increasing order.
- Min. and Max. : Correspond to the smaller and largest values in the data set.
- Range: Is the difference between the higher and lower extremes (min. and max.)

To make the study system-specific, we compressed all species within each system into a single value for each statistical descriptor, providing a comprehensive representation of the each system. To develop this protocol, we considered five different IPA systems (Mo, W, V, Nb, and Ta) and two HPA systems (PMo and AsMo), focusing on predicting the intercept parameter of the linear scaling. For each system, we computed the aforementioned statistical descriptors across the various speciation models. We then consolidated all species into a median value, resulting in a single value

for each statistical descriptor per system. This methodology is summarized in Figure 3.9.

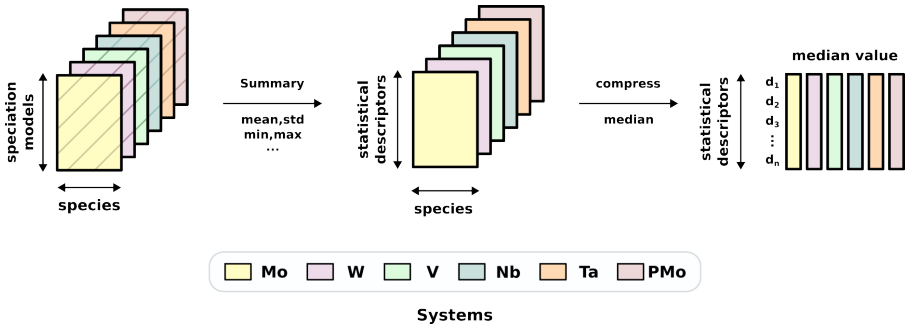

Figure 3.9: Schematical depiction of the formation constants transformation to statistical descriptors.

In parallel, we computed the linear regression of each system individually to extract the slope and intercept parameters. Initially, for each of the selected descriptors a linear regression was performed to correlate that descriptor to the proper intercept value. The results of the linear regressions indicated that a more complex algorithm was needed. Due to the scarcity of data, too complex Machine Learning (ML) algorithms could not be used and avoiding overfitting was a complicated job. For this reason among the multiple regression models, the Multi-Linear Regression (MLR) model was chosen. Overfitting could be defined as the production of an analysis that corresponds too closely or exactly to a particular set of data. Therefore, it may fail to extrapolate data outside the training set. In other words, training a model with more features than data points must be avoided.

From all the calculated statistical descriptors, all the combinations of 2 to 4 descriptors were generated. Combinations of more than four descriptors were discarded to avoid overfitting. In parallel, the scaling parameters

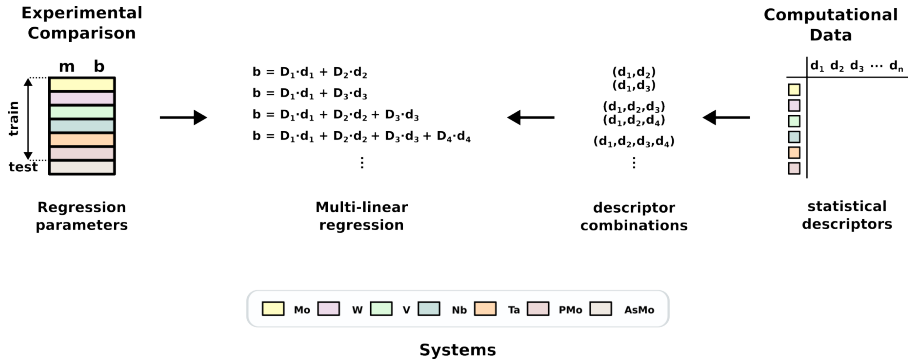

Figure 3.10: Scheme from the generation of the Multi-Linear Regression model, using calculated regression parameters to train the model, and the statistical descriptors as input parameters of the model.

(slope and intercept) for each of the systems were calculated with a linear regression between the median values across the set of speciation models and the experimental formation constants. The five IPAs (Mo,W, V,Nb and Ta) and the HPA (PMo) were chosen as training set, and the HPA (AsMo) system was selected for the test set. An scheme of the treatment of the calculated formation constants is depicted in Figure 3.10. After applying the MLR model, the different combinations of features were sorted according to the absolute error. The ten combinations with lower error are shown in Table 3.2.

As aforementioned, avoiding overfitting is vital, for this reason, as far as possible only combinations of three features are selected. In this sense, two combinations are highlighted above the rest due to only containing three features but with similar results to combinations of 4 features. To balance the difficulty of obtaining more data (full characterization of new molecular sets with available experimental data with later application of the POMSimulator methodology), we proceeded with a cross-validation strat-

Table 3.2: Best combination of features in the Multi-Linear Regression model sorted by decreasin r^2 scoring value. b_{pred} value represent the intercept parameter prediction for the AsMo system, while the $|b_{err}|$ value represent the absolute value of the difference between the predicted value and the calculated one.

Features	N_{feat}	r^2	b_{pred}	$\|b_{err}\|$
(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**	**0.70**
(std, Q_3, max, range)	4	0.967	-7.29	0.70
(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. Q_3

egy. Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data, in our case we employed a leave-two-out validation.

For every combination of features we selected two systems for the validation process, while the other five were used as training set (all 5 IPAs and the 2 HPA are used in this section). For each combination of features, there were twenty-one different combinations of training and test sets. The results obtained after applying the aforementioned cross-validation are depicted in a heat map in Figure 3.11. A heatmap is a 2-dimensional data visualization technique that represents the magnitude of individual values within a dataset as a color. The different combinations are set up in the vertical axis, while the different systems are organised in the horizontal axis. Color scale and numeric values correspond to the difference between the median b value for every instance where a system is part of the test set, and the actual values from individual regressions.

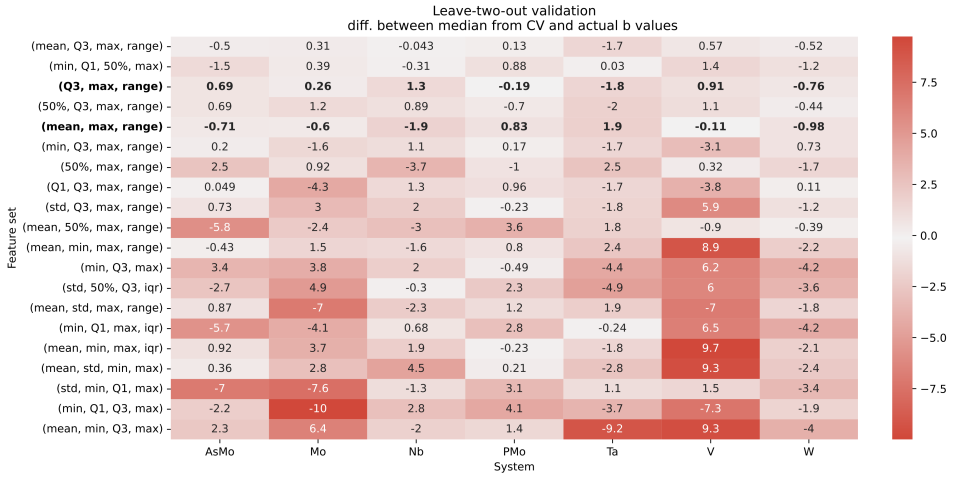

Figure 3.11: Heatmap representation of the leave-two-out validation. Y-axis corresponds to the feature sets, and X-axis to the POM systems included in the study. For each system in the X-axis the average error of considering all the possible combinations of the rest of systems is depicted. Each "pixel" is colored according to the colorbar.

Through this cross-validation approach, we could confirm the appropriateness of the previously selected combinations of features, that show a general low error in the prediction of the intercept parameter. To keep consistency we proceeded with the combination $\{Q_3$, max, range$\}$ as it was the combination with the lowest r^2 value in Table 3.2. An interesting trend that was observed during the cross validation was that when the HPAs were no used to train the model, the prediction of the HPAs intercept parameter was constantly wrong, showcasing that this data-driven model is partially aware of the chemistry behind polyoxometalates. Not considering hetero-POMs in the training resulted in high error, while introducing any of the HPA in the training set, results were closer to the proper values. Once the MLR equations for various combinations of features were obtained, the results were analyzed for the feature sets (Q3, max, range) and (mean,

max, range), which exhibited the lowest R2 values. Figure 3.12 shows the dispersion of the calculated intercept parameters.

For all the combinations of leave-two-out, the calculation of the intercept parameter is plotted as an 'x'. The exact value of the intercept parameter calculated for each system is represented as a solid circle. For both the PMo and AsMo systems, there is a value that is clearly an outlier highlighted by an empty circle. These two outliers correspond to leaving the HPA systems out of the training set, thus demonstrating the chemistry behind the model.

This proceeding was applied for the two chosen combinations of statistical descriptors.

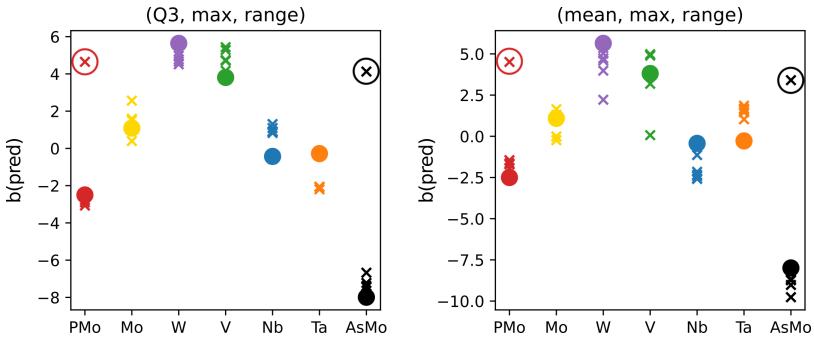

Figure 3.12: Dispersion plot for the calculation of the intercept parameter in the leave-two-out validation process. Crosses highlighted with a circle represent the outliers.

After acknowledging the correctness of the multi-regression model for predicting the intercept value of the scaling equation for formation constants, we can report the following equations, not considering the AsMo systems used as test set:

$$K_{scaled} = 0.29 \cdot K_{DFT} - b \tag{3.1}$$

$$b = 0.195 \cdot Q_3(K_{DFT}) - 0.216 \cdot max(K_{DFT}) + 0.070 \cdot range(K_{DFT}) + 12.20 \tag{3.2}$$

The overall method for the proposed universal scaling for POMs is showcased in Figure 3.13. The complete approach was applied in the AsMo

system, and the results will be presented in *Chapter 5*.

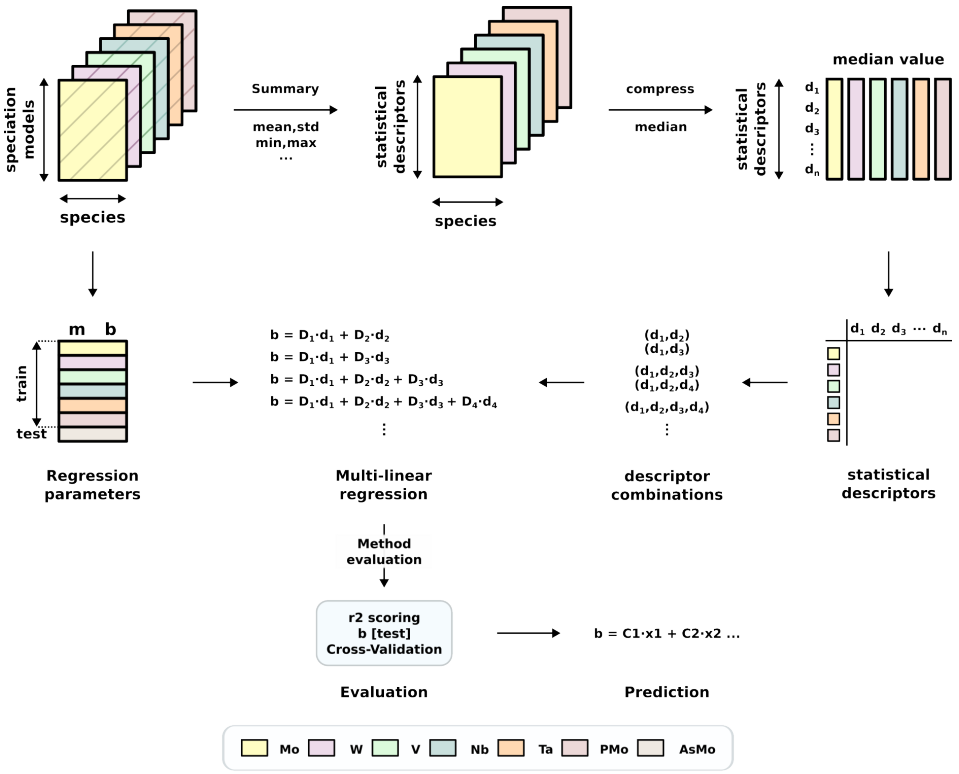

Figure 3.13: Graphical scheme of the general universal scaling methodology. At the top section of the scheme, transformation of the constants into statistical descriptors, in the middle section, generation of the Multi-Linear Regression model, and at the bottom section evaluation of the model and prediction of values.

It is important to consider, that the Multi-Linear Regression model has been trained with iso-polyoxometalates and hetero-polyoxometalates, thus to use this model with systems with different chemistry, it would be compulsory to train the model with new data related to the desired system.

Once the formation constants are properly scaled they can be used to

calculate the aqueous speciation of any desired systems. Considering the implemented changes to POMSimulator at the beginning of this chapter (random sampling, universal scaling ...), the speciation results obtained had to oversee some changes too.

3.4 Statistical analysis of Speciation Models

In order to calculate the speciation of POMs, their constants have to be properly scaled. Originally, the scaling of constants was used to select a unique speciation model, labeled as best model, and use that model to calculate the speciation diagrams, phase speciation diagrams and to study the complex reaction network. With all the implemented changes to POMSimulator, the idea of having a unique best model started fading away while the idea of having multiple speciation models being considered at the same time was arising. Even though POMSimulator was able to replicate with excellent results the speciation of 5 IPA systems using a best model, when facing more complex systems such as HPAs, a more complex solution is needed. This way, the method could become more robust against complex systems with millions of speciation models due to considering the totality of models instead of discarding them based on the scoring of a linear regression.

In this sense, the first and most simple approach was considering that the speciation of a system would be related to the totality of speciation models, in this case through the average value of the speciation diagrams of all the speciation models (*Chapter 4*). Unfortunately, the results obtained using this approach were very distant to the experimental results available. This hinted us that among the totality of models, there were

a number of them that were not predicting properly the experimental results, in addition, there were also some speciation models that were not being properly solved from a mathematical point of view. For this reason, we acknowledged that we had to find a way of classifying the speciation models according to their characteristics.

3.4.1 Clustering methods

Clustering algorithms are a type of unsupervised learning technique designed to identify multiple groupings within unlabeled data, which are known as clusters. Each cluster is formed by a set of data points that exhibit similarities based on their relationships with neighboring data points.

Utilizing clustering algorithms can be particularly crucial when dealing with unknown data. These algorithms are commonly employed to detect outliers within the dataset. Clustering algorithms can be classified into four groups according to the patterns in which the data points have to be arranged in: density models, distribution models, centroid models and hierarchical models.

- Density models: Clustering algorithms based on the density model identify regions with varying densities of data points within the data space. Data is grouped into clusters based on areas of high concentration surrounded by areas of low concentration. These clusters can take on any shape, and there are no additional constraints regarding outliers in the data space.

- Distribution models: In the distribution model, data is organized based on the probability of belonging to the same distribution. A

center point is established for a specific distribution, and as the distance of a data point from this center increases, the likelihood of it being part of that cluster decreases.

- Centroid models: The centroid model involves clustering algorithms that form clusters based on how close data points are to a central point, known as the centroid. The centroid is calculated to minimize the distance between itself and the data points. Data points are categorized based on their proximity to multiple centroids in the dataset.

- Hierarchical models: The hierarchical model, also known as connectivity-based clustering, employs a method of unsupervised machine learning that creates both top-to-bottom and bottom-up hierarchies. This approach is often applied to hierarchical data from sources such as company databases and taxonomies. While this model is more restrictive than others, it is efficient and well-suited for specific types of data clusters.

According to this classification of clustering algorithms, Table 3.3 combines one example of each category, pointing out for each method what category it is in, which parameters are required for the algorithm to run and the degree of scalability. This last property of the method is crucial due to the large number of samples (speciation models) POMSimulator generates. For this reason, K-Means was chosen as clustering algorithm due to the high size-scalability.

K-means clustering is a technique used for vector quantization, which originated in signal processing. Its primary goal is to divide n observations into k clusters. In this method, each observation is assigned to the cluster

Table 3.3: General classification of clustering algortihms, considering the type, the input parameters and the scalability of each algorithm.

Method name	**Classification**	**Parameters**	**Scalability**
K-Means	Centroids	Number of clusters	Very Large
DBSCAN	Density	Neighborhood size	Medium
Gaussian mixtures	Distribution	Many	Very Low
Birch	Hierarchical	Branching factor Threshold Optional global clusterer	Large

whose mean (known as the cluster center or cluster centroid) is closest, effectively representing a prototype for that cluster.

3.4.2 Clustering speciation models

In order to classify the speciation models, we first needed to calculate their respective speciation diagrams, and from them extract the most characteristic features. We based the characterization of speciation models on the speciation diagrams due to the easy and fast comparison with experimental information available even if only qualitative data is accessible.

In order to extract the most characteristic features out of the speciation diagram, we assumed that the concentration profiles for each of the species in the speciation diagram had a gaussian-like behaviour. Following this hypothesis, we considered some analytical peak parameters such as position, height, width and area. An schematic depiction of the featurization of the peaks is given in Figure 3.14. The height feature was defined as the maximum value of concentration across all the pH scale. The position was considered as the pH value in which the pH curve reaches the maximum value (height). The width of the peak was defined as the distance between the two points of the concentration curve at half of the max height of

the peak. The area was calculated as the numerical integration using the composite trapezoidal rule implemented in the python numpy library.

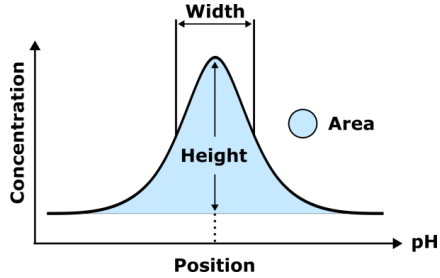

Figure 3.14: Schematical depiction of employed peak features: width, height, position and total area.

After extracting the features from the whole set of speciation models, this data is fed into the K-Means algorithm and additionally a PCA (principal components analysis) is performed to visualize the spread of the detected clusters. After that, the speciation models that have been classified in the same cluster, are plotted as the average value across them. In figure 3.15 an schematic depiction of the aforementioned process is given. To form the feature matrix, we calculated all the described features and then performed the tests with different combination of features, the most optimal to be found was the combination of height,width and position (Section 3.4.5).

After applying the PCA+K-Means algorithm clusters were formed. A chemical driven selection of groups of speciation models was applied, using any kind of data related to the studied system. Models that did not visually fit experimental data were discarded. Models that did have the same species as the experimental results but dit not agree in the relative intensities were also discarded. The selection process is totally based on a visual comparison between the clusters and the available information in

literature.

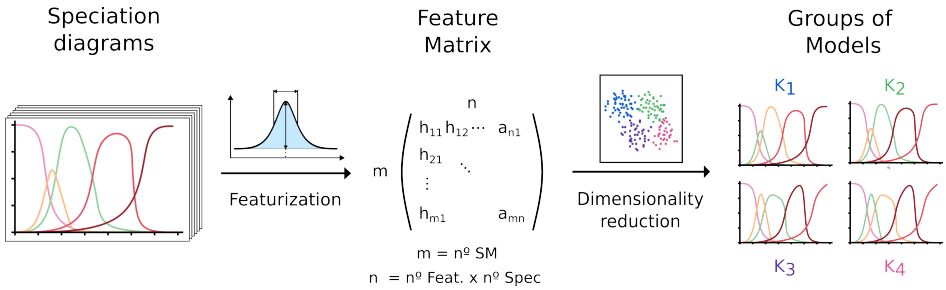

Figure 3.15: Graphical scheme of the featurization process applied to the speciation diagrams. Construction of the features matrix and application of the PCA+K-Means algorithm.

This data could be experimental speciation diagrams, syntheses conditions, stability diagrams or even spectroscopic results at a given pH. After selecting the groups of models that best replicate the considered related data, they are grouped again and if the sample is large enough, the PCA+K-Means is performed again. The second clusterization helps to further separate speciation models in order to discard possible outliers that were not found before.

In any case, after either one or two clusterizations are carried out, a further refinement of the results takes place. With all the speciation models grouped again, we select a species (usually the main species at high pH) and generate a box and whiskers plot for that species across the pH axis,considering the remaining speciation models. The speciation models that fall outside the whiskers (outliers) are discarded. To better understand this process, Figure 3.16 depicts an schematic example of the proposed methodology.

The remaining speciation models are considered the 'best group' of speciation models and the responsible for the speciation of the studied system.

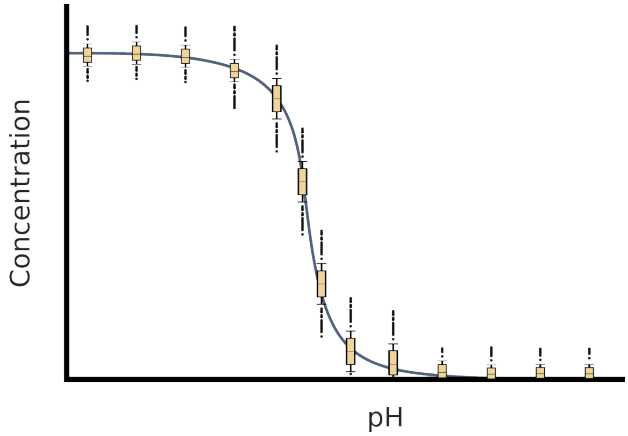

Figure 3.16: Boxplot representation of the concentration for a given species used to apply a boxplot filtering of speciation models.

With this group of models, we can then generate speciation diagrams, speciation phase diagrams and study the reaction network. It is important to mention, that for the best group of speciation models, we can calculate an error associated with the averaging of their speciation. This error is represented as an error band in the speciation diagrams. The widest bands represent that the certainty of the prediction is worse than for those whose error band is narrower.

Figures 3.15 and 3.17 represent the two halves of a global workflow, designed to select the speciation models that better reflect the available experimental information for a given system.

Through this workflow, we first consider that speciation models might be different from each other, forming cluster or groups of models that behave in a similar way. On the other hand, the second half of the workflow considers that models that have been classified in the same group of models should behave similar, for this reason the boxplot filtration is needed, to

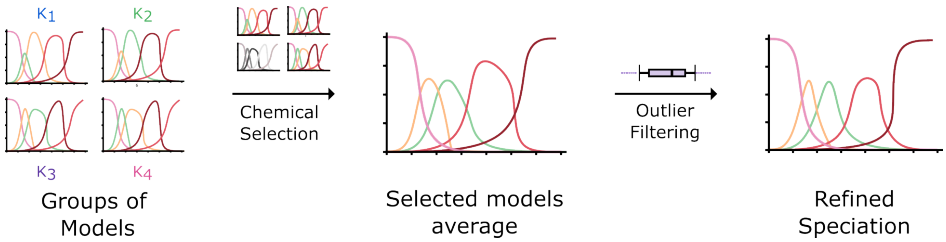

Figure 3.17: Schematical depiction of the chemical-driven selection of clusters of speciation models and outlier filtering of the selected models.

discard models that don't follow the general trend. The whole workflow is reflected in Figure 3.18, shading in different colors the two halves: in turquoise the first half, that considers variability; in purple the second half which considers the similarity inside the selection.

After applying the workflow, we end up considering speciation models in the order of tens of thousands, making the POMSimulator methodology more robust and less dependent on quantitative experimental results. Originally, without experimental formation constants, it was impossible neither to scale DFT formation constants nor choosing a best model. With this new strategy, if we are able to scale constants (which we can do if we apply the methodology described in Section 3.3), we can solve the complex speciation of polyoxometalates in a robust way.

However, before trusting the results obtained from using this new methodology, we first need to validate it by comparing the original results which were able to reproduce experimental results, with the results obtained after applying this strategy.

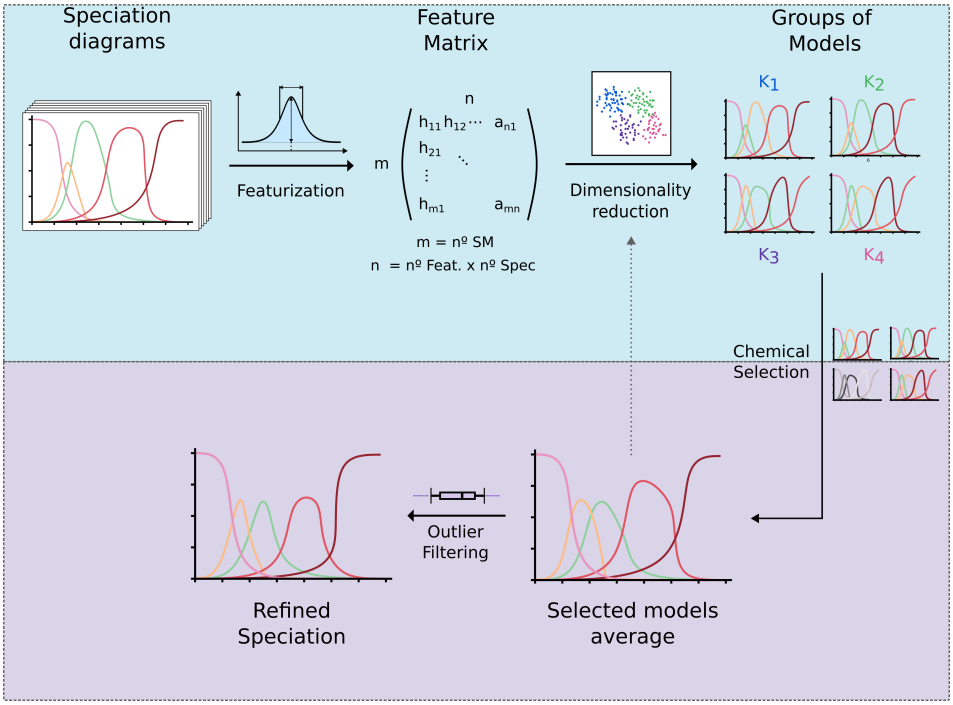

Figure 3.18: General pipeline of the statistical treatment of speciation models. On the top side, shaded in blue, the featurization and clustering process, which considers the diversity in the speciation. On the bottom side, shaded in purple, the chemical-driven selection of clusters and refinement of speciation diagrams, which considers the similarities of the grouped models.

3.4.3 Method validation: statistical pipeline

In order to make sure that the results obtained through the use of the proposed workflow, we tested it on some of the already studied systems. In that sense, the five different IPA systems that had been studied already were a perfect target to prove the new methodology. It was important that the new results were comparable to the original ones which were already reproducing the experimental speciation.

In the first place we tested the new methodology with the smaller stud-

ied system, the tungsten IPA system (W). The W system only contained 31k speciation models after removing the speciation models that were unable to be mathematically solved. After applying the featurization and clustering on the whole set of models, eight different clusters were obtained. In Figure 3.19 a depiction of the 8 clusters and the experimental speciation diagrams is given. After comparing the average speciation diagram of the eight clusters with the experimental one, Cluster 1 was chosen, as it was the one that better reproduced the experimental results[97].

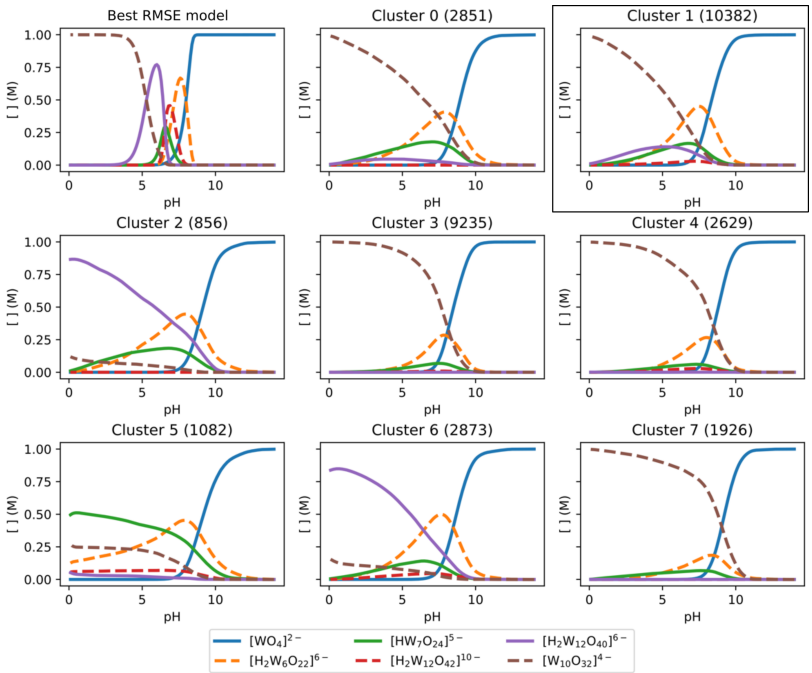

Figure 3.19: Average speciation diagrams for the generated clusters after applying the pipeling to the W IPA system. Parameters used to compute the speciation: temperature = 298.15K and C_W = 0.1M.

To further compare the two methodologies (original RMSE approach

and new statistical analysis approach), we studied how the speciation models were sorted into different clusters according to their respective RMSE value. For this reason we sorted the whole set of speciation models in ascending order of their respective RMSE value. With the sorted list of models, we divided them into 5 groups of 6k models and a last group with the remaining 1.8k models. With this classification we compared with a bar plot (3.20), in which clusters did the speciation models sort themselves. This classification showed that the best speciation models according to RMSE value, were sorted into the cluster that we had chosen (Cluster 1), demonstrating that our chemistry-driven approximation agreed with a pure mathematical RMSE-based approximation.

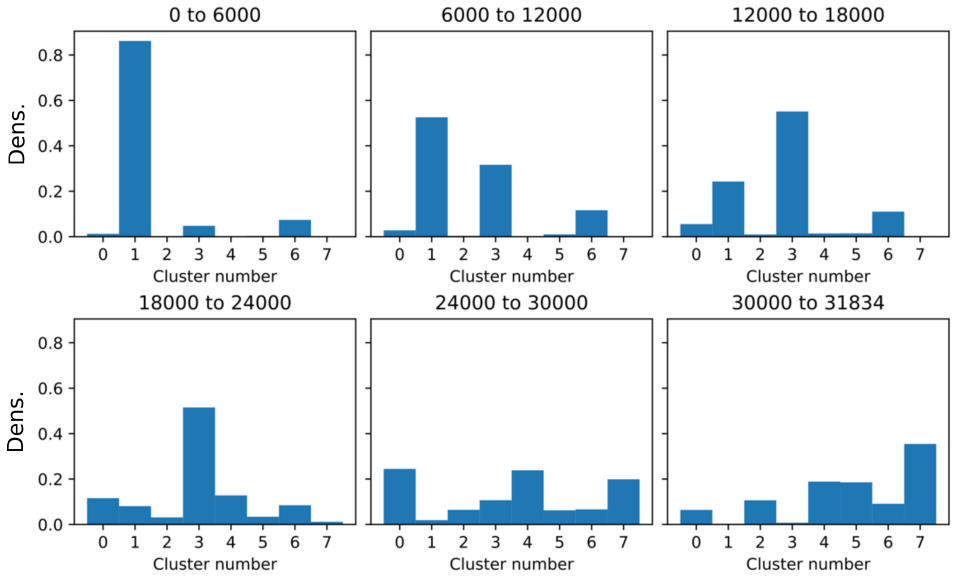

Figure 3.20: Comparison of the best model approach and the statistical treatment of speciation models. The bar plot is divided into six different plots, each of them considering 6k speciation models sorted by the RMSE value. The different clusters are organised across the horizontal axis. the vertical axis represents the density of models sorted in each cluster.

Noting that both approaches pointed to the same cluster, we proceeded with cluster 1. The next step in the workflow was the filtering of outliers. We selected the $\{W_{10}\}$ species to filter the outliers as it was the predominant species in acid pH. After removing the undesired speciation models, the size of the group decreased to 1k speciation models, which was around 10% of the initial selected models. As a result of applying the whole pipeline to the W system, we obtained comparable results with our original results (3.21).

The two main species, the $\{W_{10}\}$ in acid media and the $\{WO_4\}$ in basic pH, were perfectly reproduced. The intermediate species were not reproduced as accurately as the previous two, but the positions of their maximums and the relative position between them were in agreement with previous results. If the calculated the error band is considered after averaging the speciation models, the relative intensities of the peaks that we are not perfectly reproducing could be slightly different, much closer to the experimental and previous results.

Figure 3.21 depicts the speciation diagrams of different stages of the new methodology as well as the experimental reference used[97]. Overall, the new methodology was able to reproduce the speciation of the W system with a good agreement with previous results[73]. It was important that the new methodology was robust and scalable so it could be used in far more large and complex systems as the HPA phospho-molybdate system. Thus, following with the evaluation of the new approach, we considered the niobium system as it was larger and more complex than the W system. The Nb system was composed of 136k valid speciation models. For this reason, after the features were extracted, two consecutive clusterizations were performed. To better classify the speciation models, eleven groups

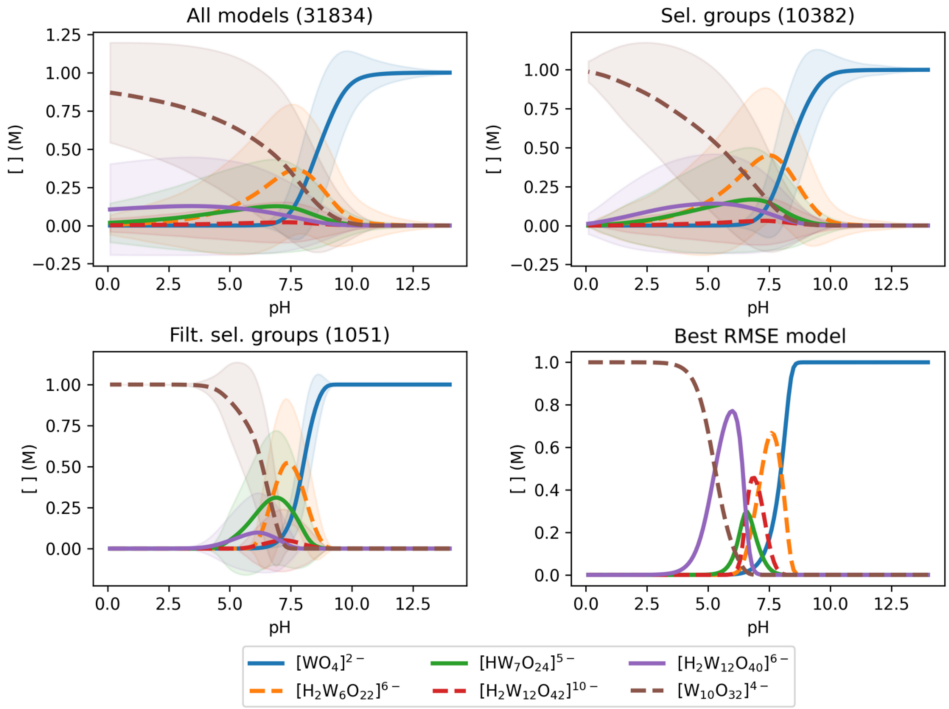

Figure 3.21: General application of the statistical treatment of speciation models to the W IPA system. Top left panel considers the totality of speciation models, top right panel onlly considers the selected speciation models while the bottom left panel considers only the models that remain after discarding outlier models. Bottom right panel represent the experimental[97] speciation the W system. Parameters used to compute the speciation: temperature = 298.15K and C_W = 0.1M.

were considered for the first clustering step. The first one was aimed to selecting the models that best predicted the main species at acid $\{Nb_{10}\}$ and basic $\{Nb_6\}$ pH. After selecting the desired clusters, the second clusterization was performed, which was aimed at predicting the intermediate species.

For this reason, only the groups with good agreement were selected. In this case, due to the good accordance of the average speciation of the

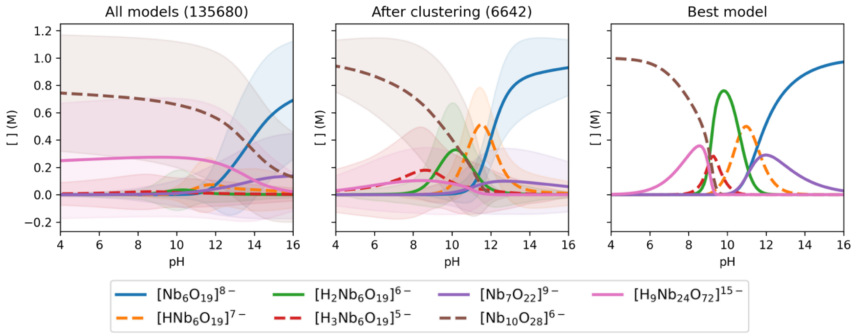

Figure 3.22: Application of the statistical treatment to the Nb IPA system. From left to right, all possible speciation models, models selected after applying the clusterization and speciation diagram of the speciation models selected as "best" according to its RMSE value. Parameters used to compute the speciation: temperature = 298.15K and $C_{Nb}W = 0.1$M.

selected models, no further refinement was needed. The results obtained after applying the new methodology to the Nb system are depicted in Figure 3.22. In this Figure, the results obtained after applying the workflow are compared with the results obtained in the original best model approach. In this case, the new methodology can reproduce with an excellent agreement the previous results, as all the peaks obtained appear in the exact positions and the relative position between them are also identical. Moreover, the intensities considering the error band fit the results previously obtained.

Following the validation process and the robustness of the method, we studied the V system, which was the largest and most complex system ever studied with POMSimulator, with the new methodology. The results of applying the statistical pipeline to the V system are depicted in Figure 3.23. The good agreement between the outcome of POMSimulator and the experimental results[98], proof that the new methodology is quite robust and that it can be applied to really complex systems, where the speciation consists of a large set of species like the vanadium or the HPA-PMo system.

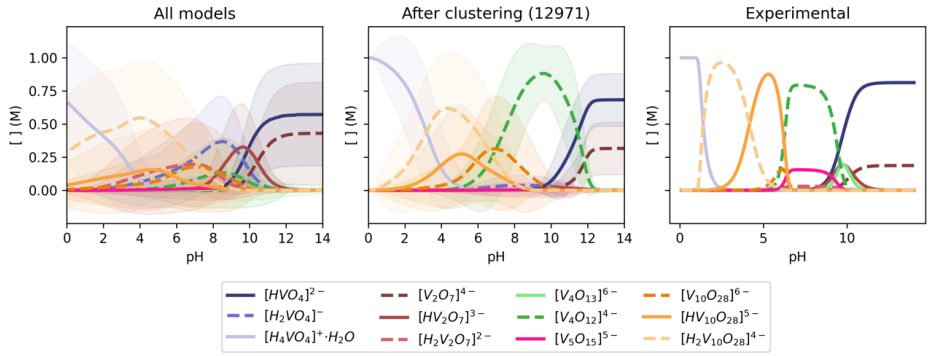

Figure 3.23: Proposed speciation of the vanadium system using the clustering approach, and comparison with experimental results[98]. Parameters used to compute the speciation: temperature = 298.15K and C_V = 0.1M.

3.4.4 Method validation: random sampling

After validating the new methodology for the Nb and V system, we considered the Nb system to test if the random sampling could affect the results on the speciation. To do so, we considered a 10% of the 135k speciation models and applied the new pipeline to those models. The results (Figure 3.24) showed and excellent agreement between the subsample final speciation and the speciation of the best model. These results were also comparable with the ones obtained in the full set of speciation models obtained in Figure 3.22.

Due to the randomness of the sampling process, we needed to verify whether the seed could also influence the results. To address this, we repeated the process using different seeds for random sampling and applied the new approach to each sample set. In this instance, we performed a single clustering and selected only one cluster. The key objective of this validation step was not necessarily to achieve results that matched the experimental

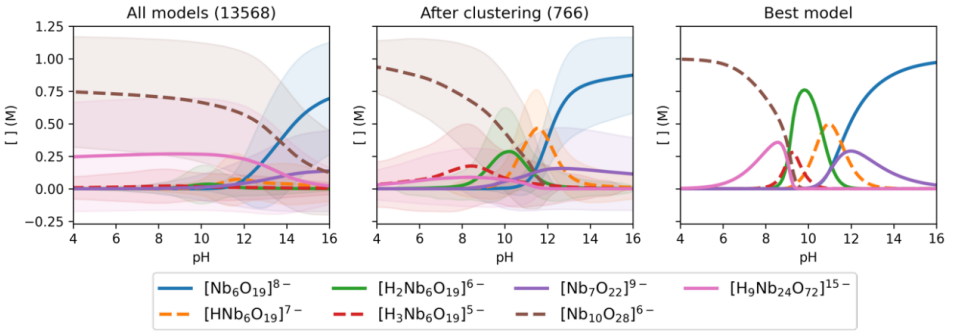

Figure 3.24: Application of the statistical treatment of speciation models to a sample of the Nb system. Parameters used to compute the speciation: temperature = 298.15K and C_{Nb} = 0.1M.

data or the best speciation model, but rather to ensure consistency in the speciation across different seeds. The results of this step of the validation process are depicted in Figure 3.25.

These results indicate that the seed does not have much effect on the outcome. It is true that the lower left and lower middle speciation diagrams are slightly different from the rest, the overall result is similar.

3.4.5 Method validation: featurization

Another step that needed to be validated before the whole new methodology was tested on a new system, was the feature selection. Previously in Section 3.4, it had been mentioned that for each concentration peak, the height, the width, the position and the area were selected as features. To test how many of these features were strictly necessary we considered the W system and applied each possible combination of features for a total of 15 (Figure 3.26).

From all these combinations, there are some (width, width + area or width + pos + area) that were considerably inaccurate, failing to predict the

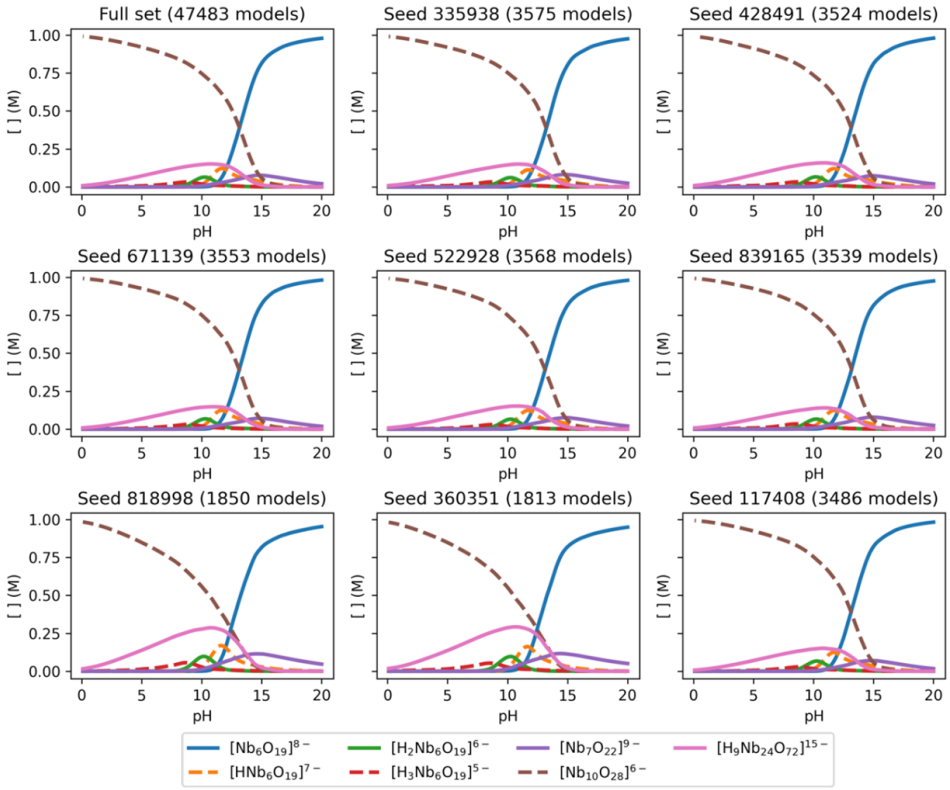

Figure 3.25: Comparison of the application of the statistical treatment to different samples of the Nb system, each of them generated using different seeds. Parameters used to compute the speciation: temperature = 298.15K and C_{Nb} = 0.1M.

speciation. While other combinations were quite more accurate, the combination of width, position and height was the one with better agreement. That being said, depending on the studied system, the feature selection could be slightly different, without compromising the utilization of the new methodology.

The new methodology was applied to the phospho-molybdate and arseno-

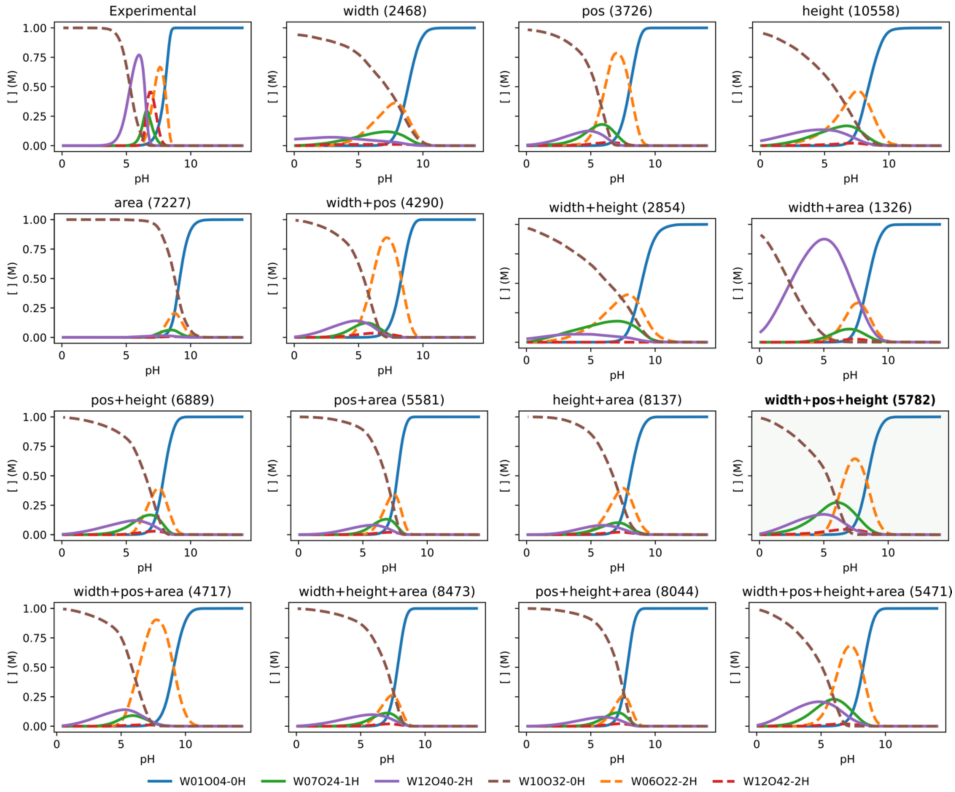

Figure 3.26: Comparison of the outcome of applying the statistical treatment to the W system, considering different combinations of features in the peak featurisation process. Parameters used to compute the speciation: temperature = 298.15K and C_W = 0.1M.

molybdate system. The results of each of these systems can be found in *Chapter 4 and Chapter 5* respectively.

3.5 Conclusions

In this chapter we have adapted the POMSimulator framework to extend its applicability from isopolyoxometalates to more complex heteropolyoxometalates. In order to deal with the increasing complexity, we improved the resolution of speciation models and achieved a 20-fold speed-up, enhancing the scalability of the POMSimulator method to large and intricate systems like the HPA. The increase in resolution speed paves the way for larger-scale studies for different polyoxometalate systems.

We have developed a new universal scaling methodology for polyoxometalates which remarkably is independent of the DFT method employed to characterize the molecular set. The new methodology builds on the preliminary results hinting to a constant slope parameter for the linear scaling of formation constants. In contrast, the non-constant intercept parameter can be predicted using a Multi-Linear Regression model capable of discerning between systems with different chemistry properties such as the IPAs and the HPAs. This approach enables us to obtain accurate results even in absence of experimental formation constants which were previously compulsory to have.

Through this chapter we have also optimized a chemical-driven statistics pipeline capable of classifying speciation models. The new clustering methodology allows the POMSimulator framework to work on large sets of data, increasing the robustness of the obtained results. Using this approach we can also determine the error associated with our predictions, having a suitable metric to assess their quality. Using both methodologies at the same time has enabled POMSimulator to evolve to the next level of data processing, in line with current efforts in data-driven science.

Chapter 4

Phosphomolybdates

4.1 Introduction

Molybdenum based heteropolyoxometalates have been very well known for the past 200 years since the discovery of the phospho-molybdic acid by Berzelius in 1826[1]. The phospho-molybdate Keggin anion and its lacunary derivatives have centered the attention of many of the studies in the field due to its properties and versatility[99–101]. After studying the most important IPA systems (Mo, W, V, Nb and Ta) with POMSimulator[72–74], we aimed to apply the methodology for understanding the formation of HPA systems, so we started with the famous Keggin anion. The phosphomolybdate (PMo) system was selected for that purpose, as there is plenty of experimental data about its aqueous speciation and formation constants are also provided in the literature[39,102].

The first step in POMSimulator's workflow is to define the molecular set of building blocks. The DFT characterization is an important step as

the reaction network generated for the system will be more or less complex depending on the molecules that are included in the set, and the interconnectivity of those molecules.

4.1.1 Molecular set

To generate the molecular set for the PMo system, we identified the main species in the results provided by the group of Prof. E. Cadot[39]. They reported a speciation diagram of the PMo system in which the deconstruction of the Keggin anion while pH was increasing was described. From that speciation diagram we recognized that some species to be included in the set. Those species go from monomeric structures to complex structures like the Keggin anion and the Keggin lacunary. In total 49 species were considered in the molecular set divided in sixteen different nuclearities.

From smaller to larger molecules, we first started with the molybdic acid and phosphoric acid (Figure 4.1). The molybdenum monomer was considered in three protonation states, from 0 protons to the 2 protons form.

Figure 4.1: Left: the molybdate anion $[MoO_4]^{2-}$ structure. Right: the phosphoric acid H_3PO_4 structure. Blue spheres correspond to the Mo atoms, red spheres correspond to O atoms and yellow-orange spheres correspond to P atoms.

On the other hand, the phosphorus monomer structure was studied in all the protonation states of the phosphoric acid, from 0 to 3 protons. It was important that we considered all the protonations of the phosphoric acid because the respective pKa of the acid are 2.14, 7.20 and 12.37 and the pH range studied in POMSimulator goes from 0 to 14 to model the acid-base chemistry of water.

After the monomeric structures, the simplest molecule was the $\{P_2Mo_5\}$ which corresponds to the PMo Strandberg structure[103]. This structure consists of a ring of five molybdenum atoms chained by bridging oxygen atoms. On the top and bottom of that five-membered ring a phosphate anion is placed. The full structure resembles the discus used in athleticism. On the other hand, in Figure 4.2, on the left side of the figure, there is depicted the $\{PMo_3\}$ structure which is believed to take an important role in the formation of larger and more complex structures. For instance, the $\{XM_3\}$ motif is repeated four times in the Keggin structure, sharing the X atom between the four trimers.

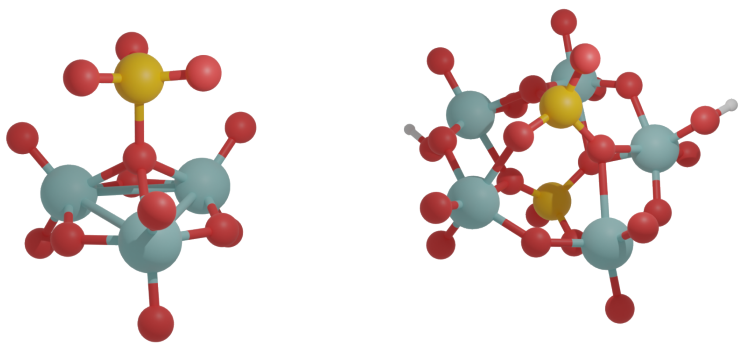

Figure 4.2: Left: the $\{PMo_3\}$ structure. Right: the Strandberg anion $\{P_2Mo_5\}$ structure. Blue spheres correspond to the Mo atoms, red spheres correspond to O atoms and yellow-orange spheres correspond to P atoms.

The next two important molecules in the set of the phospho-molybdate system are the $\{PMo_9\}$ trilacunary structures (Figure 4.3). Both structures are related by a hydration reaction involving three water molecules increasing the number of oxygen atoms from 31 to 34. In the experimental speciation diagram, two stuctures are proposed for the $\{Mo_9\}$ structure, labeled as $\{A-Mo_9\}$ and $\{B-Mo_9\}$ (orange lines). In relation to the aforementioned trimeric structure, the two $\{PMo_9\}$ isomers are formed by a central $\{XO_4\}$ atom with the three trimers surrounding it. These structures are also called Keggin tri-lacunary, and they can be capped with other metals to obtain materials with different properties[101,104].

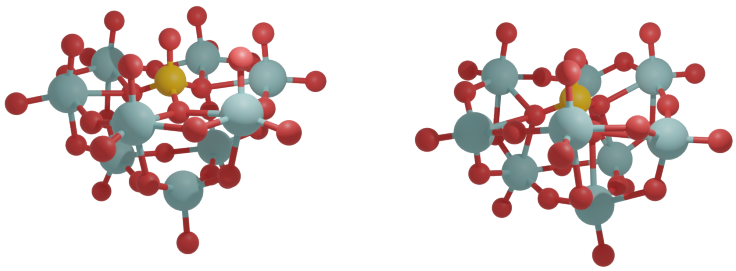

Figure 4.3: $\{PMo_9\}$ isomeric structures. The left structure corresponds to the $\{PMo_9O_{34}\}$ and the right one to $\{PMo_9O_{31}\}$. Blue spheres correspond to the Mo atoms, red spheres correspond to O atoms and yellow-orange spheres correspond to P atoms.

Last, the Keggin[105] and Keggin lacunary[99,100] structures were essential parts of the molecular set (Figure 4.4). These two structures are the most important in the PMo system due to their abundant presence in the experimental speciation results at low pH and to their extensive applications in different fields.

In the molecular set of the PMo system, for the Keggin structures we have only considered the α isomer, which is the most stable one[106]. All the isomeric forms of the Keggin structure are depicted as polyhedral rep-

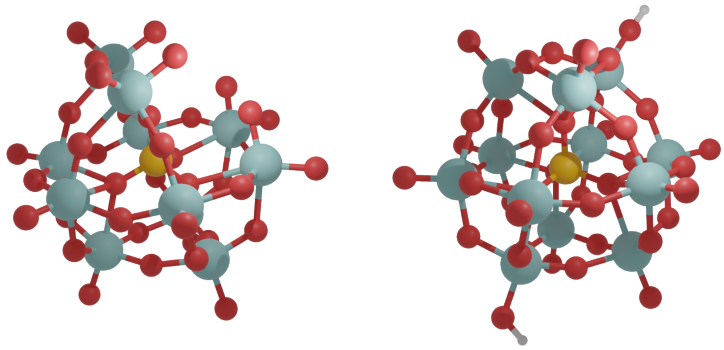

Figure 4.4: On the left part of the panel, the Keggin lacunary structure $\{PMo_{11}\}$. On the right side, the Keggin anion structure $\{PMo_{12}\}$.Blue spheres correspond to the Mo atoms, red spheres correspond to O atoms and yellow-orange spheres correspond to P atoms.

resentation in Figure 4.5. From the α to ε isomers, the difference lies in the orientation of the trimeric substructures within the main framework. In each isomerization one of the trimers is rotated by 60 degrees, and at each isomerization a different one is rotated. This way, the α isomer has the trimers joined by the vertices and the ε isomer has its trimers joined by the edges.

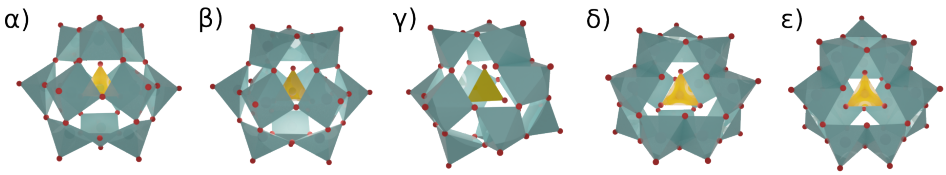

Figure 4.5: From left to right, the different Keggin structural isomers named from α to ε. At each stage, a different trimer is rotated by 60 degrees. Blue polyhedrons correspond to Mo and yellow-orange polyhedrons correspond to P.

While the species previously described are the ones present in the speciation diagrams, we can't forget the intermediate species. These molecules might not seem particularly important, however they are crucial to form

large molecules like the Keggin anion. Across the molecular set for the PMo system, we considered three different trimer species ($\{Mo_3O_9\}$, $\{Mo_3O_{10}\}$ and $\{Mo_3O_{11}\}$) and two different dimers ($\{Mo_2O_7\}$ and $\{Mo_2O_8\}$).

The molecular set can be accessed through the following reference [107] in the ioChem-BD platform[108,109].

4.1.2 Chemical Reaction Network

The complexity of the HPA systems is far beyond the IPA systems. In the PMo case, the 49 species of the molecular set are related by a total of 109 different chemical reactions. Among these reactions there are acid-base, condensation, addition, and hydration reactions. The totality of reactions generate a convoluted and largely interconnected reaction network, which translates into a massive amount of speciation models ($\approx$ 300 million). This reaction network is shown in Figure 4.6 as a 3D plot of the different species as nodes (full circles), and the reactions represented as colored lines based on the Gibbs free energy of the reaction. The position of the species is calculated according to their stoichiometry:

- The horizontal axis position depends mostly on the number of phosphorus atoms to which it is added the ratio between the number of oxygen atoms and molybdenum atoms: $P + O/Mo$.
- The vertical axis position depends on the amount of molybdenum atoms.
- The depth axis position depends on the number of hydrogen atoms

Based on this positioning, the small molecules in the set can be found in the bottom left corner of Figure 4.6. In the bottom center part of the

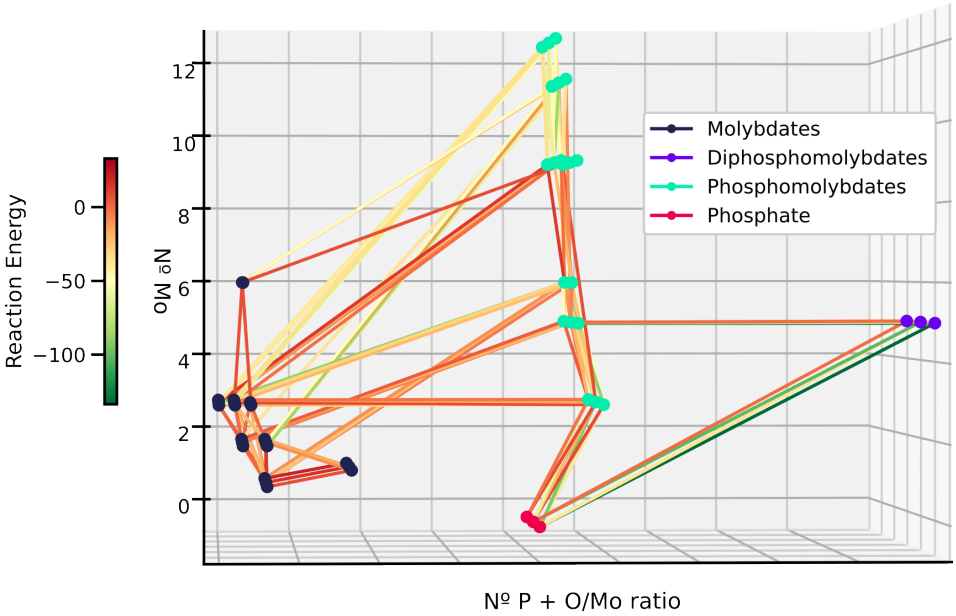

Figure 4.6: Complete Chemical Reaction Network for the PMo system. It is formed by 49 molecules and 109 reactions (acid-base, condensation, addition and hydration). The color of the lines represent the energy of the reaction. Nodes (molecules) are colored according to the composition of the molecule.

panel the phosphate monomer species is found. In the center of the panel, more complex molecules containing both molybdenum and phosphorus are positioned. The species containing two phosphorus atoms fall on the right side of the figure.

The complexity of the CRN (high interconnectivity) makes it unfeasible to obtain a simple reaction mechanism out of it. For this reason, it is imperative to simplify the CRN into something reasonable to work with. In this sense, the aforesaid statistical treatment of speciation models appears as a solution to reduce the problem of complexity.

4.2 Speciation results

The Chemical Reaction Network for the PMo system we have just defined is composed of around 300 millions speciation models. Due to hardware limitations it was not possible to solve the totality of the system. In *Chapter 3*, we applied a random sampling of the speciation models and obtained excellent results, so consequently we randomly sampled 1% of the speciation models of the PMo system. Once we had solved the 3 million (1% of 300M) speciation models, we applied two different methodologies to obtain the speciation of the PMo system. On the one hand we applied the original methodology (POMSimulator), which consisted of performing millions of linear regressions and selecting the speciation model with the lowest RMSE value from those regression. On the other hand we applied the methodology described in *Chapter 3* (POMSimulator 2.0), which consisted of applying a clustering algorithm (K-Means) to the speciation models and describing the speciation as an average of the selected models.

In all cases, the speciation outcome was compared to the speciation diagram obtained by the group of Prof. E. Cadot[39] (Figure 4.7). The experimental speciation diagram contains three main species in acid pH, which are the Keggin $\{PMo_{12}\}$ anion, the Keggin lacunary $\{PMo_{11}\}$ and the Strandberg anion $\{P_2Mo_5\}$. With lower concentration we also find the two $\{A-PMo_9\}$ at pH=2 and the $\{B-PMo_9\}$ at pH=5. Finally at a $pH > 7$ we find the phosphate anion.

This diagram does not account for the protonation states of the species that it reports. For this reason, applying the POMSimulator methodology, which does consider acid-base reactions, to the phosphomolybdate system will help to discern with higher accuracy the aqueous speciation for this

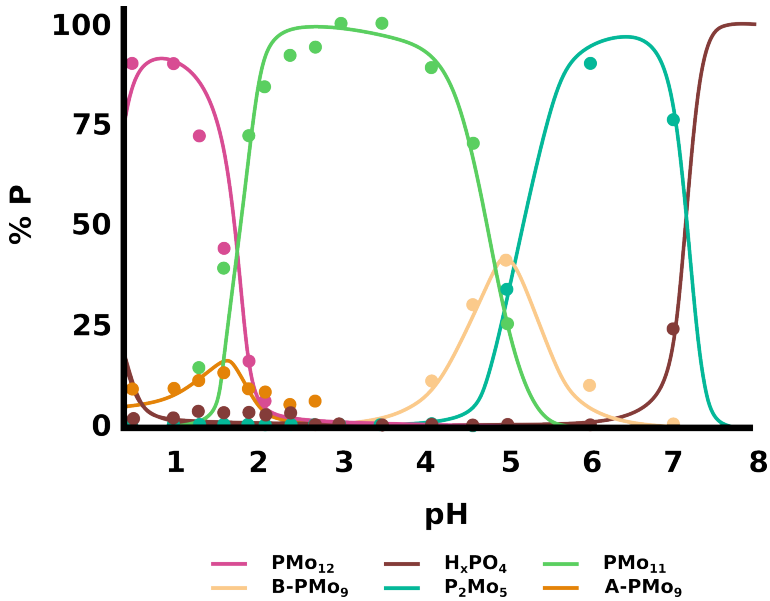

Figure 4.7: Experimental speciation diagram, extracted from [39]. The colors have been changed to match the color code used in POMSimulator.

system.

From our side, as aforementioned, we computed the speciation of the PMo system using two different approaches. In the first place, we employed the original methodology and after performing the linear regression of the 3 million calculated speciation models, we chose the three models with lowest RMSE scoring. We then employed those SMs to predict the speciation diagrams of the PMo system (Figure 4.8).

If we compare the speciation diagrams that were obtained with the experimental speciation diagram, we found that while some species are present in the same pH range, the rest do not fit in any way the experiments. Moreover, in any of the speciation diagrams the $\{P_2Mo_5\}$ species or any of

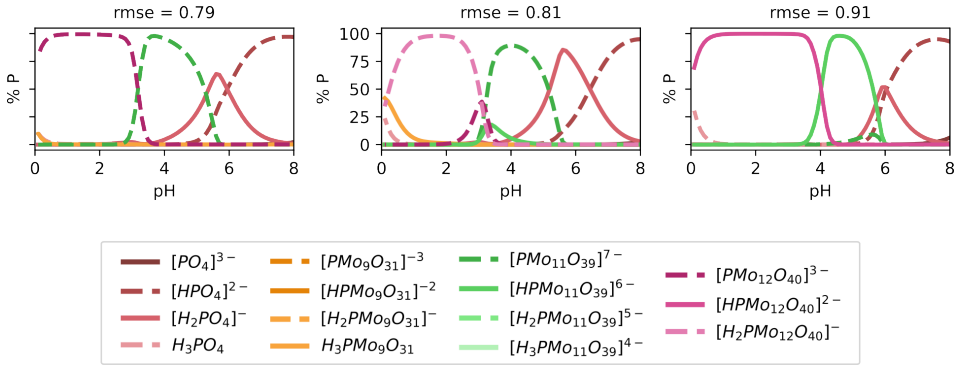

Figure 4.8: Speciation diagrams generated by POMSimulator, corresponding to the three models with the lowest rmse value from the linear scaling. Parameters used to compute the speciation: temperature = 298.15K, C_{Mo} = 0.12M and C_P = 0.01M.

their protonation states, was not present. This absence was the straw that broke the camel's back, and demonstrated the need for a new approach.

After the original POMSimulator methodology failed to properly predict the speciation of the PMo system, we posed a second approach that consisted in solving all the speciation models and averaging the speciation diagrams. This new point of view supposed a clear improvement in the robustness of POMSimulator as it was considering all the data that was available. Furthermore, this strategy unlocks the possibility of estimating the error of the prediction in the speciation, determining the standard deviations across the model set.

The average speciation diagrams are depicted in Figure 4.9, in which two different representations of the speciation diagrams are proposed. In the left panel of the figure, the average speciation is plotted according to the absolute concentration values, which considering the stoichiometric relations between the monomeric species and the larger clusters as the Keggin anion or the Keggin lacunary, leaves us with very low concentration values

of the larger cluster, which makes it difficult to observe and compare the results. In the right panel of the figure, the average speciation is represented as molar percentage of either the metal or heteroatom. This visualization of the concentrations, allows to locate the peaks corresponding to large clusters, and also to compare with the experimental speciation diagram.

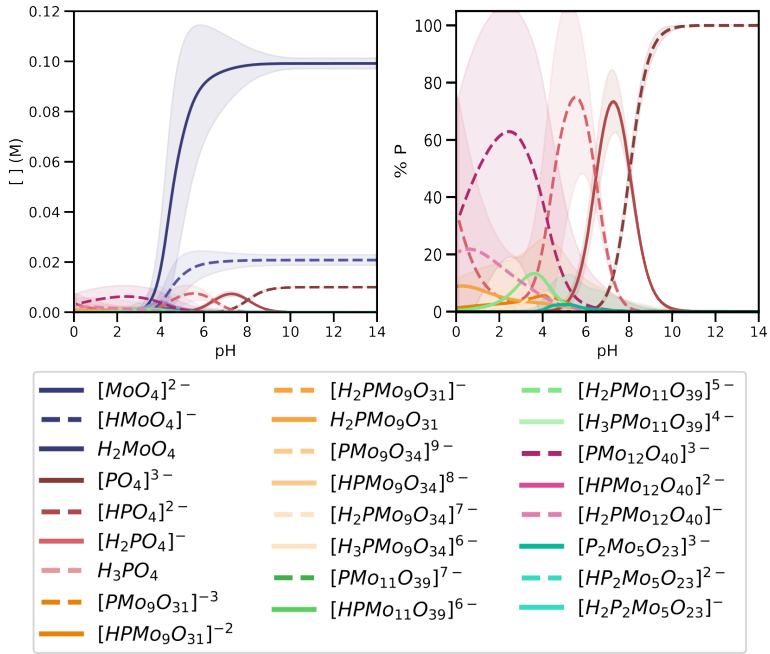

Figure 4.9: Average speciation diagram for the whole set of speciation models (1.5M), represented as raw concentrations in the left panel and as % of P in the right panel. Parameters used to compute the speciation: temperature = 298.15K, C_{Mo} = 0.12M and C_P = 0.01M.

Nonetheless, it also evidenced that among the speciation models, not all of them had the same behaviour (showing huge standar deviation) and that SMs in some way could be classified according to their speciation diagrams. At this point, the concept of classifying the speciation models emerged, for a posterior selection of the groups of models that would properly predict

the experimental results. To do so, we applied the methodology described in *Chapter 3*.

Before jumping to the clustering of the speciation models of the PMo system, we needed to scale the previously calculated formation constants. In *Chapter 3* a new methodology to scale formation constants using a Multi-Linear Regression model was described and proposed as a solution to the lack of experimental information and a way to generalise POMSimulator.

While the universal scaling methodology was not fully developed at the moment of studying the PMo set, the wide availability of experimental formation constants for this system enabled us to propose an intermediate solution. Instead of scaling each speciation model with the parameters from their linear regression, all the speciation models were scaled using the same parameters, which were obtained by averaging the parameters of all the speciation models. To clarify, at this point there were three different scaling methodologies:

1. The original POMSimulator: performing a linear regression for each speciation model and scaling each model with their own parameters.

2. Intermediate approach: performing a linear regression for each speciation model, average the regression parameters and scale all the speciation models using the same slope and intercept (used to predict the PMo speciation).

3. POMSimulator 2.0 methodology: applying the universal scaling equation to predict the intercept parameter, and scale all the speciation models using the same slope and intercept.

With this approach, we were still performing millions of linear regressions, but as we were going to average the outcome of the speciation, we were treating all the speciation models in the same manner. For the PMo system, the average scaling parameters were the following: $m = 0.28$, $b = -2.02$.

Once the formation constants were scaled, we proceeded to apply the pipeline described in the *3.4 Statistical Treatment* section in *Chapter 3*. First we applied the featurization process across nearly 1.7M speciation models, and next we applied the PCA+K-Means algorithm using the Height, Width and Position features, dividing the sample into 15 different clusters. The results from the first clusterization are depicted in Figure 4.10, in which the 15 (the number of clusters depends on the number of speciation models) clusters are represented as an average speciation diagram from the models within the group.

From all the 15 clusters, not all of them were properly predicting the speciation of the PMo system. When we compared the results of the clusters with the experimental results depicted in Figure 4.7, we selected Cluster 3 (highlighted in bold). This group of models was predicting the appearance of the Keggin anion at acid pH in the same range as the experimental speciation diagram. It was also predicting the rest of the species present in that diagram like the $\{PMo_{11}\}$ or the $\{P_2Mo_5\}$. It was important that all the species present in the experimental results were present in the selected clusters, but it was also important that the relative positions and intensities between them were also in agreement with the experimental speciation diagram.

Even though only one cluster was selected, the number of SMs was around 150k which was large enough to proceed with a second clusterization

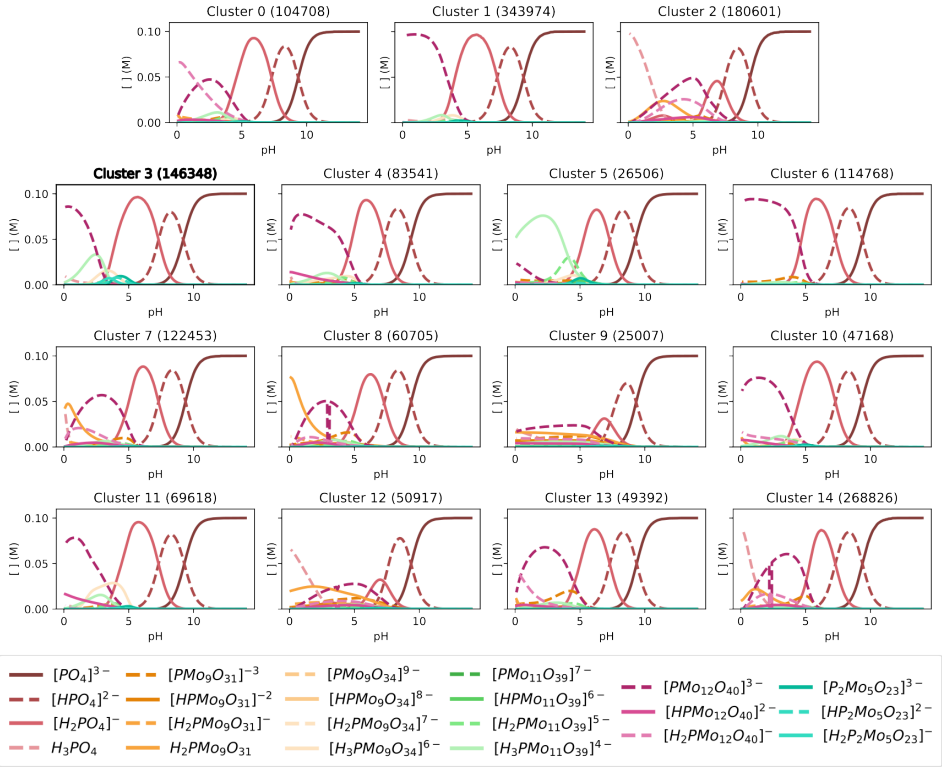

Figure 4.10: Average speciation diagram for the 15 generated clusters in the first clustering. From this clusters, the selected group (Cluster 3) is highlighted in bold. Parameters used to compute the speciation: temperature = 298.15K, $C_{Mo} = 0.12$M and $C_P = 0.01$M.

to refine our prediction.

For the second clusterization, we employed the same features we had previously calculated, but in this case we assigned 8 clusters to the K-Means algorithm to consider the decrease in the sample after the first clusterization.

The results of this stage of the pipeline are represented in Figure 4.11. In this clusterization step, we observed a clear division between models that

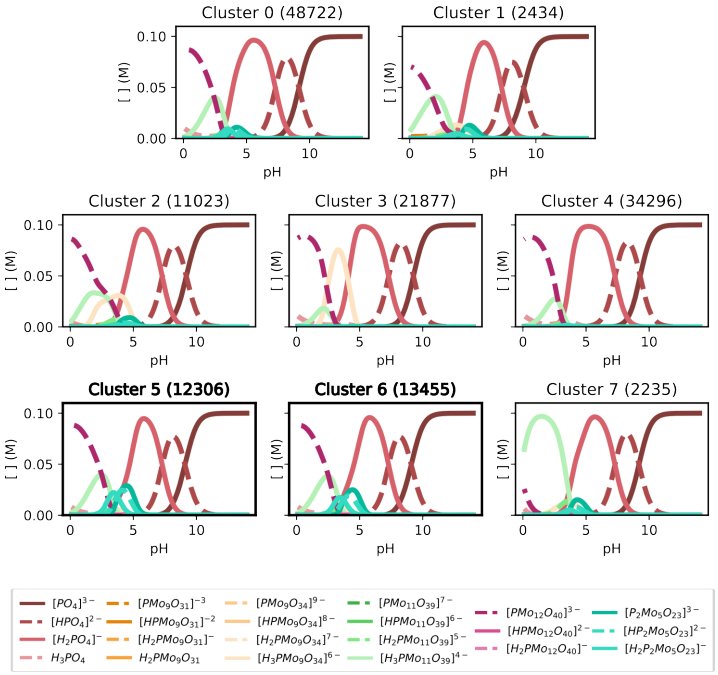

Figure 4.11: Average speciation diagrams for the clusters generated in the second clustering step. Selected clusters (5 and 6) are higlighted in bold. Parameters used to compute the speciation: temperature = 298.15K, $C_{Mo} = 0.12$M and $C_P = 0.01$M.

were under-predicting the Keggin anion concentration, and those that were under predicting the rest of species present. But among this classification, Cluster 5 and Cluster 6 were properly predicting the formation of the main species ($\{PMo_{11}\}$) and the minor species too.

Following the statistical workflow we applied a further refinement to the selected speciation models, removing the speciation models outside the general trend of the group.

After removing the outliers, we plotted the average speciation diagram of the selected SM. A new feature added with the use of average speciation

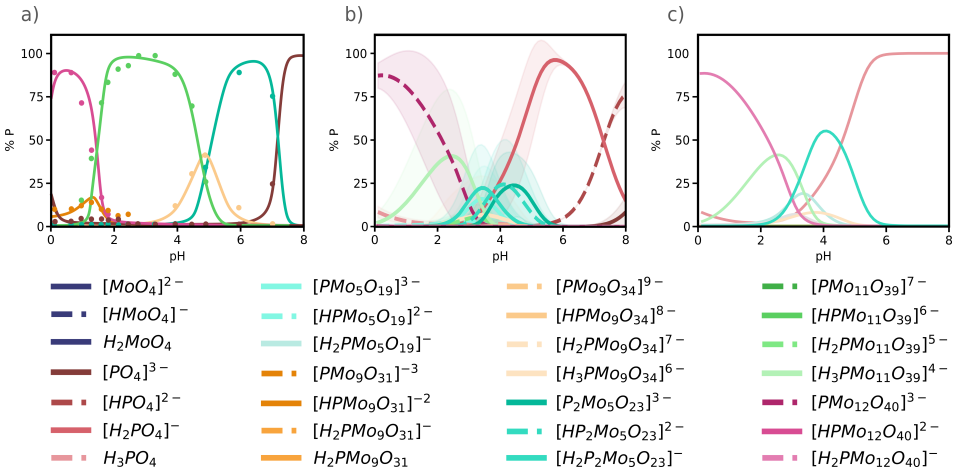

Figure 4.12: Final speciation for the PMo system. On the left panel of the figure the experimental speciation diagram, from [39]. In the center panel, the average speciation diagram with the error band of the 25k selected models. On the right panel, average speciation diagrams grouped by nuclearities, without error band. Parameters used to compute the speciation: temperature = 298.15K, $C_{Mo} = 0.12$M and $C_P = 0.01$M.

diagrams is that we were able to calculate an error in our prediction, in this case as the standard deviation (σ). This error is displayed as a shaded band which goes from the concentration (C_i) at a given point to $C_i \pm \sigma$. Experimental results do not consider different protonation states, so for the sake of comparison we converted our speciation diagram of the PMo system to a speciation diagram that only considered the nuclearities and not the individual species. To do so, we summed the contributions of the different protonation states of the same nuclearity. In this representation, we did not consider the error band to better visualize the speciation. These two representations (species and nuclearities) are depicted in Figure 4.12 together with the experimental speciation diagram.

In the center of the figure, we find the speciation diagram by species and if we compare it to the experimental one in the left side of the Figure,

what we observe is that all the species present are present in the prediction, agreeing with the relative positions. On the other hand, the intensities of the peaks does not match perfectly the experimental diagram. However, if we consider that our prediction has an associated error, we could reproduce the intensities of the experimental peaks.

An important contribution from our study is that our prediction gives light to a problem in the experimental speciation diagram. In Figure 4.7, the experimental speciation assigns one of the peaks to $\{A - PMo_9\}$ and another to $\{B - PMo_9\}$. From our side, we propose that these assignments might be different. The $\{A - PMo_9\}$ peak corresponds to the $\{PMo_9O_{31}\}$ species while the $\{B - PMo_9\}$ peak corresponds to the $\{PMo_9O_{34}\}$ species. In addition, we were able to predict the presence of the $\{PMo_5\}$, an intermediate species which had never been reported before in the literature, and that is crucial for the formation of the Strandberg anion.

To complement the speciation studies, we were able to compute the speciation phase diagrams for the PMo system. Speciation phase diagrams give a good overview of the general speciation of the system in a wider range of conditions (Initial metal concentration and pH). We expanded the selected 25k SM to consider models that were initially discarded, that would be suitable for the speciation in different metal ratios, ending up with 75k SM.

When dealing with HPA systems, the concentration parameter was ambiguous as it could point to the metal atom or to the heteroatom. For this reason, to consider both concentrations at the same time, we used the concentration ratio (R) between the metal and the heteroatom. So instead of visualising the speciation as in the previous applications of POMSimulator, we used the concentration ratio. Another important consideration when

dealing with speciation phase diagrams is that only the main species at a given R and pH point is represented. Thus, to build the phase diagrams, we solve the speciation models at each ratio value and at each pH point, and extract the main species. This way, we are building a matrix with m rows and n columns, where m represent the amount of ratios considered and n represents the amount of pH points. The higher these values are, a better resolution is obtained, but it also implies a higher computational cost.

Another important consideration was the way the data is visualized. Through this chapter, all the speciation diagrams are represented as the % of P, but in the PMo system, we should also acknowledge the major form in which Mo is present. As the experimental results were always reported as % P, we did not visualize from the point of view of Mo, but as speciation phase diagrams haven't been reported for HPAs in literature, we considered studying from both points of view (Figure 4.13).

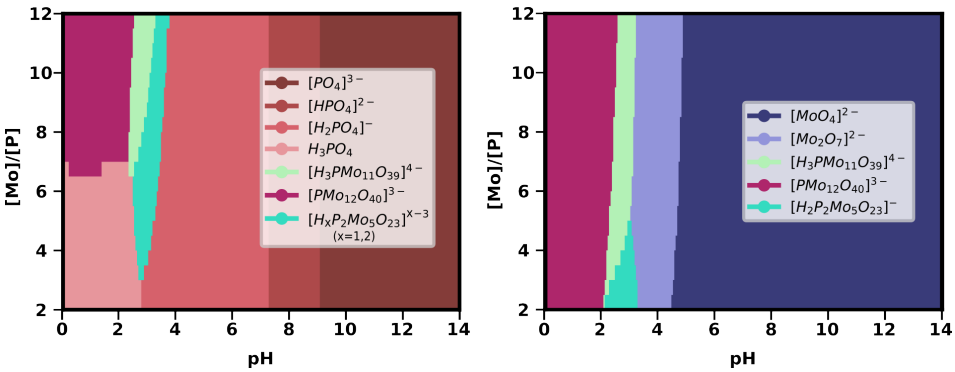

Figure 4.13: Speciation phase diagrams for 75k speciation models. The left panel represents the % of P and the right panel represents the % of Mo. Parameters used to compute the speciation: temperature = 298.15K, C_{Mo} ranges from 0.12M to 0.02M and C_P = 0.01M.

Having both perspective of the same speciation phase diagram allows us

to make a more complete reading of the results. At low pH values, we can see that all the molybdenum present in solution is contained in the Keggin form, but at low concentration ratios we can find free phosphoric acid in the aqueous solution. If pH is slightly increased, we can see the transformation of the Keggin anion into the Keggin lacunary, but at low concentration ratios, the Strandberg anion appears as the main species. From the point of view of molybdenum distribution, the $\{Mo_2\}$ structure is present at pH values comprised between 3 and 5, to transform into free molybdate anion at higher pH values. From the point of view of the phosphorus, we can appreciate the releasing of free phosphoric acid from the Strandberg anion at higher pH values. This free phosphoric acid has it own acid-base speciation.

It is noteworthy that in these phase diagrams, we lose information about non-majority species like the $\{PMo_9\}$ isomers and the $\{PMo_5\}$ structure. But in contrast we gain a general overview of the system and what to expect in a wider range of conditions.

It is worth noting that there are no reported experimental speciation phase diagrams for the PMo system or any other HPA system. Our contribution on predicting speciation phase diagrams, is significant because it provides experimental researchers with a valuable tool to optimize their synthetic pathways. By offering a general map of reactant ratios and pH conditions, this predictive model can help guide the controlled synthesis of polyoxometalates, making it a key resource for advancing experimental work in the field.

4.3 Mechanism insights

After successfully applying the clustering methodology to the PMo system, we ended up with 25k speciation models with a proper speciation. For this reason, continuing with the idea of considering all the available data, we decided to reduce the complex reaction network presented at the beginning of this chapter (4.6)

As it has been explained in *Chapter 2* and *Chapter 3*, each speciation model consists of a set of chemical reactions ranging from acid-base to addition reactions. To discern a probable formation mechanism of the Keggin anion, we studied the reactions of all the 25k speciation models in order to extract any possible trend, or any repeated reaction across all SMs. All acid-base reactions were discarded in this part of the analysis as they are constant in all SMs.

With the remaining reactions, we studied for each nuclearity all the reactions that formed it, and then selected that with higher frequency. It is noteworthy that in most cases, the most repeated reaction were not the most thermodynamically favourable. Acid-base reactions play a crucial role in the stability and reactivity of each nuclearity After selecting the reaction that appears most frequently across all speciation models for each nuclearity (except for the MoO_4 and PO_4), we organised the molecules in a similar as in Figure 4.6. This way, the species that didn't contain any P atom were gathered on the left side, while the species with one P atom were positioned in the center. The Strandberg anion, containing two P atoms was positioned on the right side. In the vertical axis, nuclearities were positioned according to the number of Mo atoms, leaving the small species at the bottom of the figure, and the large clusters at the top. A

depiction of the Reaction Network of the 75k models is given in Figure 4.14.

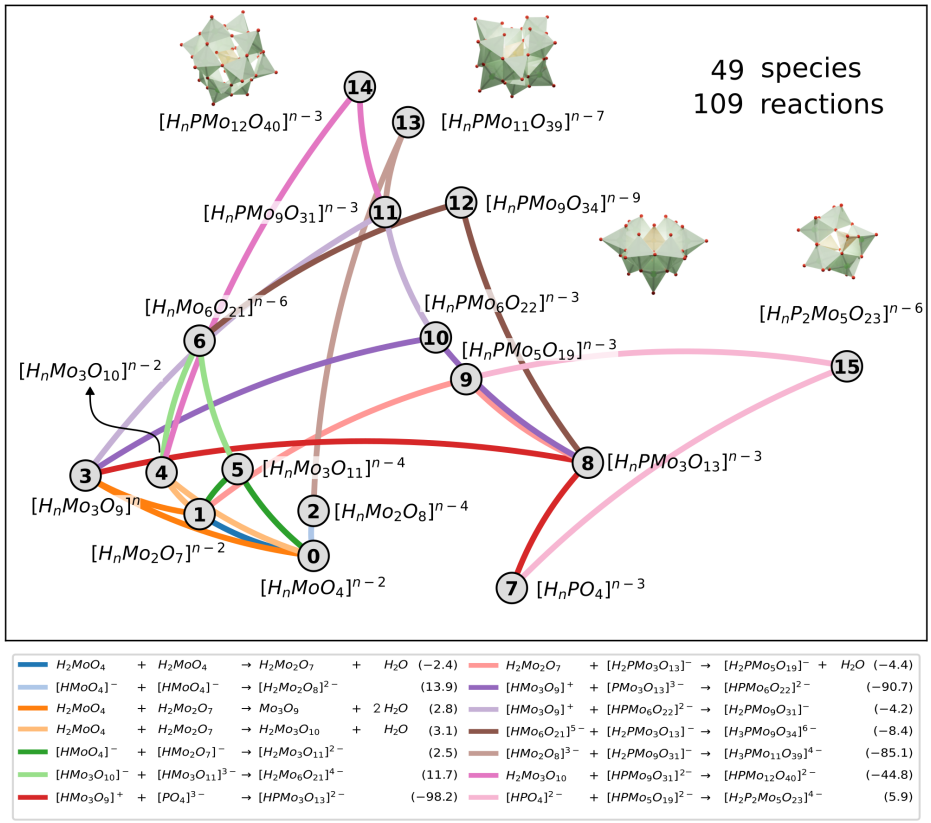

Figure 4.14: Chemical Reaction Network of the most frequent reactions in the selected 25k speciation models from the second clustering step.

From this CRN, what we can extract that the small nuclearities like the dimers and the trimers are extensively interconnected, showing the reason why these species cannot be observed in the speciation diagrams. In this sense, the main species from the speciation diagram ($\{PMo_{12}\}$, $\{PMo_{11}\}$ and $\{P_2Mo_5\}$) do not participate in any reaction as a reagent, but as a

product in all cases. The $\{PMo_5\}$ and the $\{PMo_9O_{31}\}$ nuclearites on the other hand, have a role in nucleation reactions for other species and also appear in the speciation diagrams, eventhough they are minority species.

Focusing on the formation of the Keggin anion, it has been proposed that the trimer structures play a crucial role[100,105]. Following with this idea, we can observe how the trimeric species (3, 4, 5 in Figure 4.14) take part in the nucleation reactions of the main species. For instance, the $\{Mo_3O_9\}$ participates in the formation of the $\{PMo_3O_{13}\}$ by an addition reaction with a phosphate in different protonation states. This new nuclearity follows up with an addition reaction with another $\{Mo_3O_9\}$ to form the $\{PMo_6\}$, which can suffer a new addition reaction with a trimer to form the $\{PMo_9O_{31}\}$. This cascade of reactions ends with a last addition reaction of a linear trimeric species $\{Mo_3O_{10}\}$ to finally form the Keggin anion. If the $\{PMo_9\}$ reacts with a dimer it forms the Keggin lacunary instead.

Alternatively, if the $\{PMo_3\}$ species reacts with a dimer through a condensation reaction it forms the $\{PMo_5\}$ species, which is an intermediate to form the $\{P_2Mo_5\}$ after a new addition reaction with a phosphate.

In the legend of the figure, the selected reaction are given, with the Gibbs free energy (in $\text{kcal} \cdot \text{mol}^{-1}$) associated. From these values we conclude that most reactions are exergonic and that the ones that aren't, have relatively low energies. If we consider the acid-base reactions, that we previously discarded, the energies get even more stabilized, displacing this complex multi-species equilibria to the formation of large clusters in a reaction cascade. This showcases the importance of including acid-base equilibrium into the POMSimulator methodology.

4.4 Conclusions

In this chapter we have successfully applied the methodology presented in *Chapter 3* to predict the aqueous speciation of the phosphomolybdate system.

We have uncovered the speciation phase diagram for the PMo system, reporting for the first time a phase diagram for this kind of system, using the [Mo]/[P] ratio vs pH. We have provided two different perspectives for those diagrams, either focusing on the distribution of the phosphorus species or the molybdenum-based ones, giving a general overview on the self-assembly of phosphomolybdates.

Additionally, our predicted speciation diagrams are in very good agreement with experimental data, highlighting the abundant presence of the Keggin anion $\{PMo_{12}\}$ at low pH values. The speciation diagrams that we predict also reveal the presence of the Keggin lacunary $\{PMo_{11}\}$ and the Strandberg $\{P_2Mo_5\}$ species in different pH and concentration conditions. Remarkably, the presence of the experimentally reported Strandberg anion could only be predicted after applying the statistical workflow on the phosphomolybdate system.

We have used the statistical pipeline to transform the complex reaction network for the PMo system (109 reactions) into a manageable CRN, that contains only the most frequent reactions for each nuclearity, emphasizing the importance of the trimeric species as key intermediate building blocks in the speciation of the system.

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Chapter 5

Arsenomolybdates

5.1 Introduction

In the previous chapter we successfully predicted the speciation of the phosphomolybdate system, ruled by the formation of the Keggin anion, through the statistical treatment of speciation models. For this chapter, our main goal was to use for the first time the Universal Scaling methodology outlined in *Chapter 3* and validate it targeting a new heteropolyanionic system. For this reason, we selected the arsenomolybdate system which was first reported in 1826 by Berzelius, who reported the formation of yellow species after mixing an excess of molybdic acid with arsenic acid. The AsMo system was later studied by potentiometric and spectrophotometric methods by Pettersson et al. in 1975[92]. In the beginning of the 20^{th} century the AsMo system was extensively studied and a huge amount of crystal structures were detected.

At the moment of the study of Pettersson, the composition of solid

phases was already known, but the speciation was still yet to be further investigated. Their study gave light to the conditions of the equilibrium of the AsMo system, revealing the aqueous speciation of the system. They found the formation of arseno and di-arseno molybdate species of 5 $\{As_xMo_5\}$ to 6 $\{As_xMo_6\}$ Mo atoms in low metal/heteroatom ratios, while in high ratios the presence of Keggin trilacunary species ($\{AsMo_9\}$) was detected. More recently, computational studies[93] to characterize arsenic heteropolyoxometalates, increase the importance of studying the AsMo system. In that work they establish the thermodynamics of formation of different coordination geometries with various heteroatoms, the As among them.

To apply POMSimulator to the AsMo system, we first needed to define a proper molecular set. As it was done with the PMo system, we looked at the experimental information available to see which were the most important species in solution .

5.1.1 Molecular set

The most simple molecules to be mentioned are the monomeric species. The molybdate anion $[MoO_4]^{2-}$ and the arsenate anion $[AsO_4]^{3-}$ have tetrahedral structures. These species, or their protonated forms, are used as initial reagents to form the rest of the AsMo species. The structure of both molecules is depicted in Figure 5.1.

While the As atom remains in a tetrahedral environment, the molybdenum atoms can be found in both tetrahedral and octahedral environments. An example of both coordination forms is shown in Figure 5.2, where a trimeric building block is formed by tetrahedral Mo atoms, and the Strandberg anion is formed by octahedral Mo atoms. The AsMo Strand-

Figure 5.1: Left: $[MoO_4]^{2-}$ structure. Right: $[PO_4]^{3-}$ structure. Blue: Mo atoms, the violet: As atoms and the red: O atoms.

berg structure is equivalent to the one in the PMo system, and like in the PMo system, it is formed at low metal/heteroatom ratios.

In the AsMo system the Anderson-type ($\{AsMo_6\}$) polyoxometalate evolves into the $\{As_2Mo_6\}$ species[94,95], which is very present in the aqueous speciation at various concentration ratios. The structures of the $\{As_xMo_6\}$,

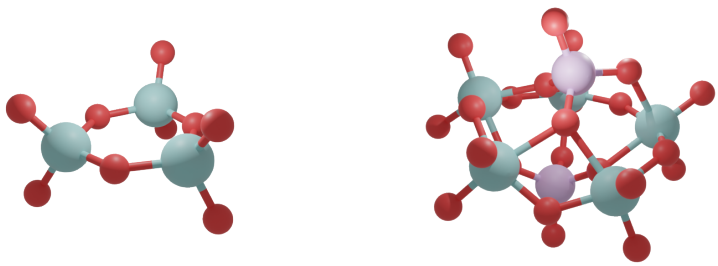

Figure 5.2: Left: Molecular structure of the Mo_3O_9 building block. Right: Molecular structure of the Strandberg anion ($As_2Mo_5O_{23}$). Blue: Mo atoms, the violet: As atoms and the red: O atoms.

where x can be one or two, are depicted in Figure 5.3. Both structures are formed by a ring of six molybdenum atoms grouped in three pairs of Mo_2O_2,

and capped by one or two tetrahedral AsO_4 units. The $\{As_2Mo_6\}$ structure is highly symmetrical (D3d point group). Similarly to the Keggin anion in the PMo system, studied in *Chapter 4*, it was important to properly predict the formation of the $\{As_2Mo_6\}$ in the speciation of the AsMo system.

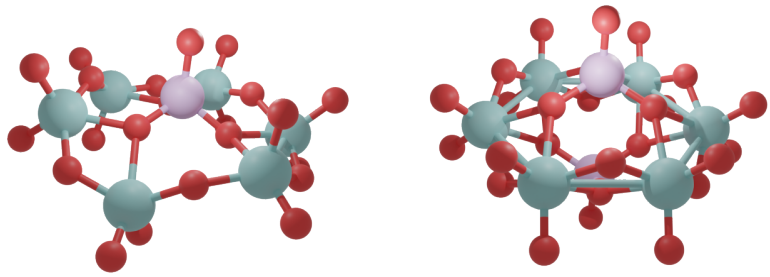

Figure 5.3: Left: AsMo Anderson-type anion ($\{AsMo_6\}$). Right: Molecular structure of $\{As_2Mo_6\}$.Blue: Mo atoms, the violet: As atoms and the red: O atoms.

In the AsMo system, the largest molecules are the Keggin tri-lacunary $\{AsMo_9\}$ anions. Like in the PMo system, both structures are related by a hydration reaction involving three water molecules. For the hydrated $\{AsMo_9\}$ species, higher protonation states are considered to counter the high negative charges. The molecular structure of these anions is shown in Figure 5.4.

Opposite to the PMo system, neither the Keggin nor the Keggin lacunary anions are found in the AsMo system experimental speciation[96]. For this reason, none of them was considered in the molecular set, for the sake of simplifying the already complex reaction network. On the other hand, like in the PMo system, smaller building blocks had to be considered to build the most complex structures in the system. Different dimer and trimer (linear and cyclic) species were considered. Intermediate building blocks like the $\{AsMo_3\}$ or the $\{Mo_6\}$ species were also taken into account.

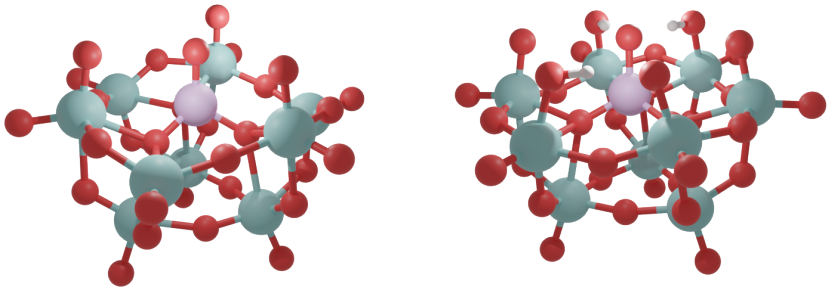

Figure 5.4: Left: $\{AsMo_9O_{31}\}$ molecular structure. Right: $\{AsMo_9O_{34}\}$ structure. Blue: Mo atoms, the violet: As atoms and the red: O atoms.

The molecular set can be accessed through the following reference [110] in the ioChem-BD platform[108,109].

5.2 Speciation results

With the molecular set properly described and fully characterized, we generated the CRN for the system, which accounted for 44 species and 96 reactions. The complete CRN of this system is depicted in Figure 5.5, with the nuclearities classified according to their stoichiometry. From bottom to top, the positions of the molecules depend on the amount of Mo atoms. From left to right, the molecules are sorted by the number of As atoms, and to separate molecules with the same number of As and Mo atoms, their horizontal position is modified by adding a fraction of the O/Mo ratio. This way we can se how the $\{Mo_3\}$ nuclearities are slightly separated in the horizontal axis. For the $\{AsMo_9\}$ nuclearities, the same ratio is applied to distinguish between the two species.

When the combinatorial equation is applied, the number of speciation models grows up to 38 million. Similarly to the PMo system, a random

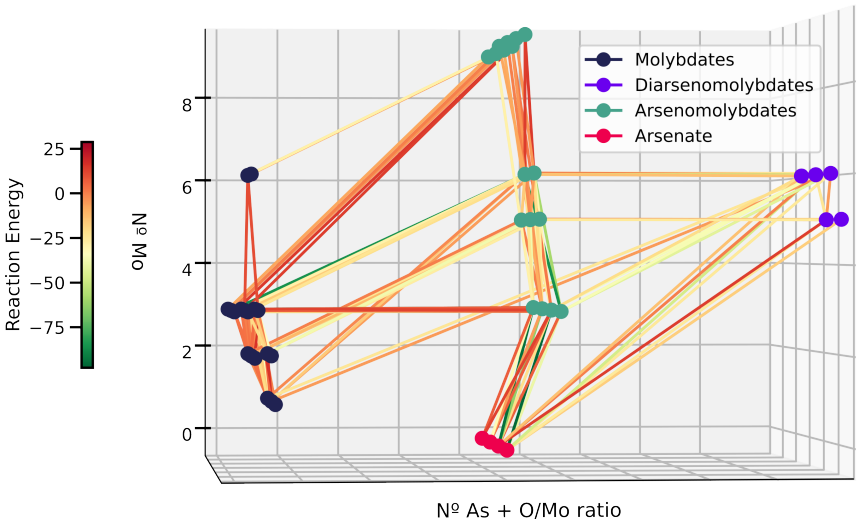

Figure 5.5: Chemical Reaction Network of the AsMo system, with the 44 species and 96 reactions. Reactions are colored according to their energy, and molecules according to the legend.

sampling was applied to the speciation models. In this case, the sample was 3% of the total, which was around 1.25 million speciation models. From there, we computed the corresponding formation constants which, as discussed along the Thesis, must be scaled properly to fit experimental results.

In *Chapter 3* we calculated the median scaling parameters of different metal systems. We employed the median value of the calculated formation constants along the speciation models to make a single linear regression for every system, reducing the hundreds of thousands of regressions that were performed originally. To better visualize the spread of the DFT formation constants, we used all the calculated values of the species with experimental formation constants, to generate box and whiskers plots (see Figure 3.3).

The linear regression of these median values for the AsMo system is

depicted in Figure 5.6, obtaining that the median scaling parameters for this system were $m = 0.33$ and $b = -7.29$.

On the other hand, we could also apply the universal scaling methodology based on the MLR model,discussed and validated through *Chapter 3*. We applyed Equations 3.1 and 3.2 to determine an alternative set of scaling parameters: $m = 0.29$ and $b = -7.99$.

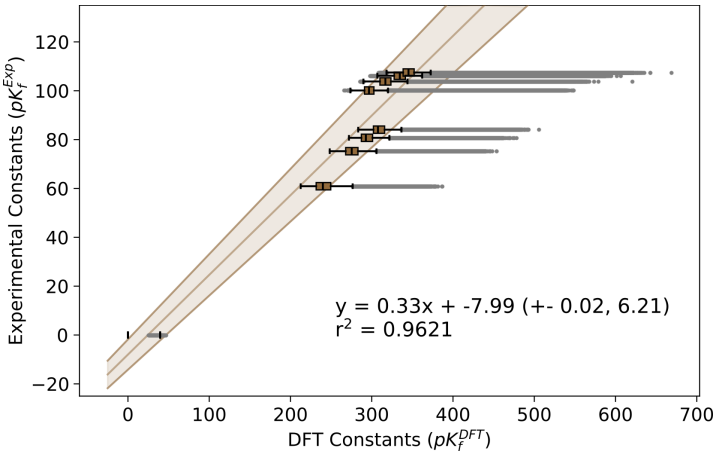

Figure 5.6: Linear regression of the experimental formation constants and the DFT formation constants represented as box and whiskers plot. The central line represents the linear regression and the outter lines correspond to the regression equation ± the corresponding error.

If we compare both outcomes, the difference in the scaling parameters is minimal and thus should not affect the speciation results. This remarks the potential of the proposed universal scaling scheme, achieving the same predictive power from the original POMSimulator methodology even in the absence of any experimental formation constants from the target system (in this case the AsMo).

To further acknowledge the effectiveness of the universal scaling methodology, we used the MLR-based scaling parameters for AsMo system to pre-

dict the speciation diagrams and speciation phase diagrams. At this point, we turned our view to the speciation diagrams reported in [92], where 5 different diagrams at varying Mo/As concentration ratios were reported.

Each of these speciation diagrams is slightly different form the previous one, showing the evolution of the system at different concentrations. From these 5 different concentration ratios, we selected four of them: 4/4, 3/1, 6/1 and 9/1. Even though the 4/4 ratio might seem the same as 1/1 ratio, we considered 40mM concentration of both molybdenum and arsenic instead of 10mM to better reproduce experimental conditions. For the rest of the ratios we employed 10mM concentration of As and multiplied that value by the selected ratio to obtain the Mo concentration.

Having a metal and a heteroatom, like in the PMo system, implied that speciation diagrams can be represented focusing on the %Mo or %As. Although the experimental diagrams were represented as the molar fraction of Mo, we can represent the same diagram from both perspectives. We followed the same clustering methodology discussed in *Chapter 3* and applied to the PMo system in *Chapter 4*.

Starting with the 4/4 ratio, the experimental speciation diagram reports the presence of three different protonation states of the key $\{As_2Mo_6\}$ anion, from pH=0 up until pH=6. It also reports the presence of the Strandberg anion at nearly neutral pH values. From pH 7 to more basic media, only monomeric $\{MoO_4\}$ species is reported. From our side (Figure 5.7), we are predicting the $\{As_2Mo_6\}$ species in the same pH range, differing in the relative intensities of the various protonation states. Nevertheless, if we consider the associated error bands, these relative intensities fit the experimental diagram. In our speciation diagrams, we also predict a second protonation state for the Strandberg anion. Considering the speciation

diagram represented as %As, we see that there is free arsenic acid as a major species.

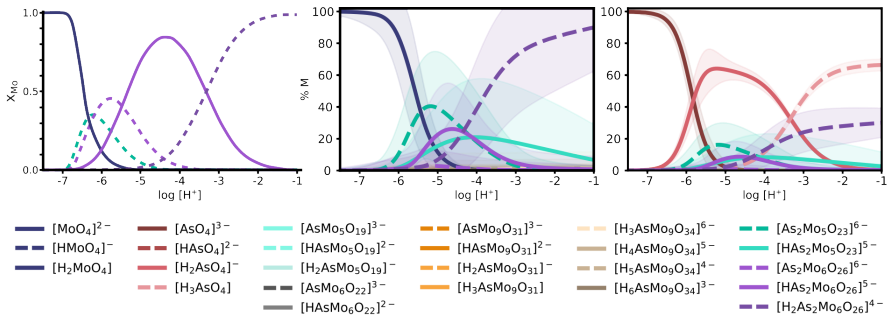

Figure 5.7: Speciation diagram for the AsMo system with a Mo/As ratio of 4/4. Left: experimental speciation diagram extracted from [92]. Center: predicted speciation diagram represented as %Mo. Right: predicted speciation diagram represented as %As. Lines are colored according to the legend. Parameters used to compute the speciation: temperature = 298.15K, $C_{Mo} = 0.04$M and $C_{As} = 0.04$M.

Moving to the 3/1 ratio (Figure 5.8), it can be appreciated that the experimental diagram is very similar to the previous one. A similar behaviour of the $\{As_2Mo_6\}$ species is reported, but with lower intensities. This decrease is explained by the presence of the $\{AsMo_9\}$ species that start to form in these concentration ratios. Again, if we look at the calculated speciation diagrams, represented as %Mo, we can observe the formation of the $\{AsMo_9\}$ species in a neutral-acid pH range. Moreover, the intensities of the rest of the peaks are slightly decreased across the acidic pH. In contrast, when looking at the %As diagram, we see that that arsenic acid is not the main species in acid media anymore, instead the $\{As_2Mo_6\}$ gains importance.

Next, the speciation corresponding to the 6/1 ratio shows a major change compared to the previous diagrams, as it is depicted in Figure 5.9. In this concentration ratios, the observed speciation trend between the

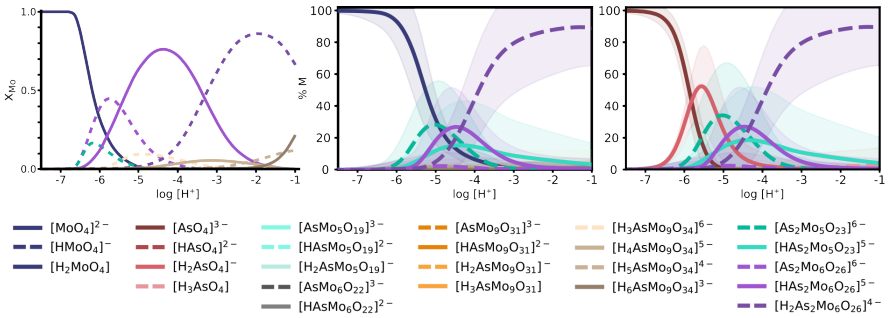

Figure 5.8: Speciation diagram for the AsMo system with a Mo/As ratio of 3/1. Left: corresponds to the experimental speciation diagram extracted from [92]. Center: predicted speciation diagram represented as %Mo. Right: predicted speciation diagram represented as %As. Lines are colored according to the legend. Parameters used to compute the speciation: temperature = 298.15K, C_{Mo} = 0.03M and C_{As} = 0.01M.

$\{As_2Mo_6\}$ and $\{AsMo_9\}$ species is reversed. The $\{H_xAsMo_9O_{34}^{x-6}\}$ species become the main species from pH=6 to pH=1, leaving the $\{As_2Mo_6\}$ as a secondary species. As in the other ratios, the Strandberg anion remains as a minor contributor at neutral pH values.

The speciation diagrams calculated with POMSimulator for this ratio show a peculiar behaviour. The MoO_4 concentration does not decrease in a sigmoidal trend as in the other cases. Below this concentration line, the Strandberg anion and the $\{As_2Mo_6\}$ species are located. In acid pH values, the $\{AsMo_9\}$ species are found in their two forms, $\{AsMo_9O_{34}\}$ and $\{AsMo_9O_{31}\}$, coexisting at the same pH ranges. In addition, the presence of and intermediate $\{AsMo_6\}$ species is predicted in a pH around 3. Opposite to the $\{MoO_4\}$, the arsenate species keep the same behaviour while decreasing its concentration.

Last, the speciation diagram corresponding to the concentration ratio 9/1 shows the total inversion of the trend (Figure 5.10). The $\{AsMo_9\}$ species appear as the main ones in acid pH, leaving the $\{As_2Mo_5\}$ and

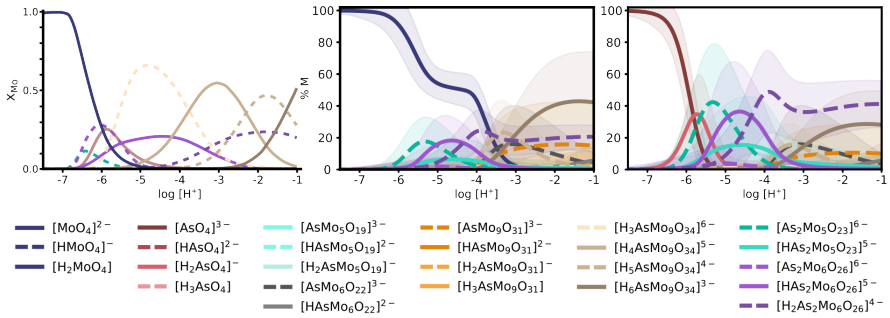

Figure 5.9: Speciation diagram for the AsMo system with a Mo/As ratio of 6/1. Left: corresponds to the experimental speciation diagram extracted from [92]. Center: predicted speciation diagram represented as %Mo. Right: predicted speciation diagram represented as %As. Lines are colored according to the legend. Parameters used to compute the speciation: temperature = 298.15K, C_{Mo} = 0.06M and C_{As} = 0.01M.

the $\{As_2Mo_6\}$ species in the background at neutral pH values. At neutral pH, the experimental speciation diagram reports the presence of a $\{Mo_7\}$ species, which was not included in the molecular set. However, due to the non-dominant character of this anion in the overall speciation, we kept this molecular set as simple as possible. The predicted speciation diagrams at this concentration ratio predict the presence of the two $\{AsMo_9\}$ species at acidic pH, while in a more neutral pH the main species remains the $\{MoO_4\}$. Underneath the molybdate concentration curve, the %Mo speciation diagram predicts the formation of the Strandberg anion as well as the $\{As_2Mo_6\}$ species. From the %As perspective, these two last molecules appear as main species, while the $\{AsMo_9\}$ species are the main ones in lower pH values.

Given the complexity of the speciation diagrams in Figures 5.7-5.10 following our previous studies, overall trends on the self-assembly of arsenomolybdates can be better represented as a speciation phase diagram. We generated the corresponding speciation phase diagram at 20 different

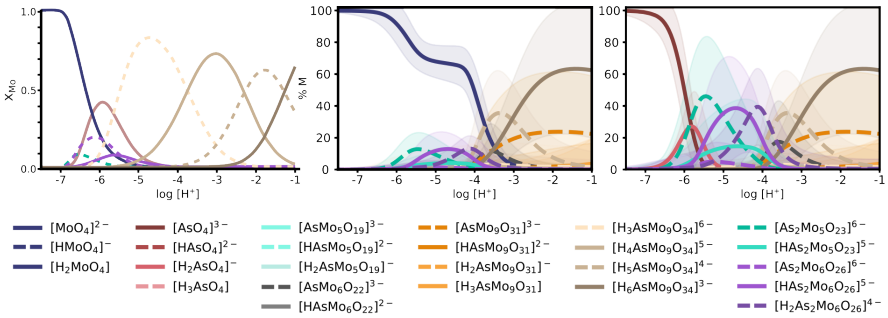

Figure 5.10: Speciation diagram for the AsMo system with a Mo/As ratio of 9/1. Left: corresponds to the experimental speciation diagram extracted from [92]. Center: predicted speciation diagram represented as %Mo. Right: predicted speciation diagram represented as %As. Lines are colored according to the legend. Parameters used to compute the speciation: temperature = 298.15K, $C_{Mo} = 0.09$M and $C_{As} = 0.01$M.

Mo/As ratios, ranging from 1/1 ratio to 9/1 and 280 pH grid points. The phase diagram was generated from 60000 speciation models, selecting the predominant species at every pair of pH and Mo/As ratio.

In the speciation diagrams shown in 5.7, the concentration ratio corresponding to the 4/4 was simulated using 40mM concentration of both Mo and As. For this particular phase diagram, the ratio corresponding to the 1/1 was calculated using 10mM concentration of both Mo and As. For this reason, the lower section of the diagram might be slightly different from the speciation diagram in Figure 5.7.

On the left side of Figure 5.11, we observe that at low pH and lower Mo/As ratios, arsenic is mainly found in the form of free arsenate (pink-red). As we increase the metal concentration, the $\{As_2Mo_6\}$ (purple) species becomes the predominant form of arsenic. At low pH and high metal concentrations, the main species is $\{AsMo_9\}$ (brownish), which aligns with the experimental speciation diagrams. When the pH increases at intermediate ratios, the $\{As_2Mo_6\}$ transforms into the Strandberg anion $\{As_2Mo_5\}$

(in turquoise), even at higher metal concentrations.

Turning to the %Mo diagram, similar trends are observed under high metal concentration and low pH conditions. However, as the ratio decreases, the $\{AsMo_9\}$ species is replaced by the $\{As_2Mo_6\}$ species. Only at intermediate acidic pH and low metal concentrations does the Strandberg anion $\{As_2Mo_5\}$ emerge as the main molybdenum species. In both cases, at higher pH levels, nucleation is not observed, and only monomeric species are present.

The success of our methodology is highlighted by the accurate characterization of the two major molecules in the acidic-pH chemistry of the arsenomolybdates: $\{As_2Mo_6\}$ and $\{AsMo_9\}$. Moreover, we also successfully predict the switch from the former to the latter as the dominant species up to pH=4 at increasing proportions of molybdenum.

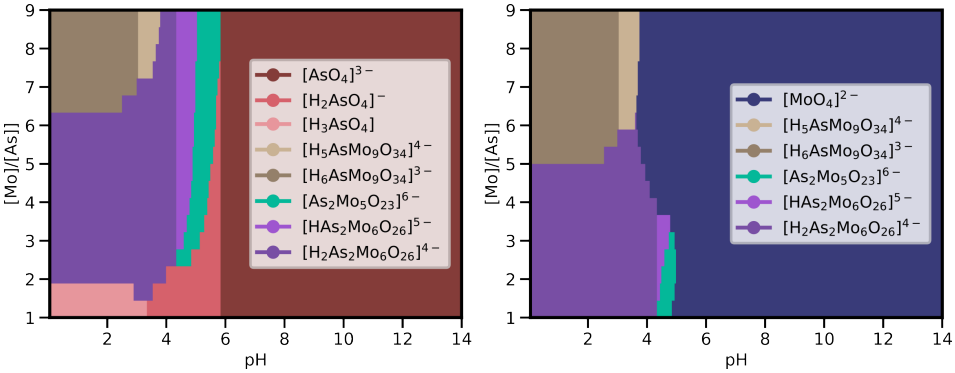

Figure 5.11: Predicted speciation phase diagram of the AsMo system. Left: represented as %As. Right: represented as %Mo. Colors are assigned according to the legend. Parameters used to compute the speciation: temperature = 298.15K, C_{Mo} ranges from 0.09M to 0.01M and C_{As} = 0.01M.

5.3 Mechanism insights

As in the previous chapter, the complex reaction network with all the reactions (Figure 5.5) was refined to only include the most frequent reactions in the speciation models selected from the last clustering step. This way the number of reactions was reduced from 96 reactions to only one reaction per nuclearity. In Figure 5.12 the CRN is depicted as a 2D plot. Acid-base reactions are omitted in this CRN as they are constant in all speciation models. That being said, the reactions described in the legend of Figure 5.12 do consider different protonation states. In the CRN, the nuclearities are sorted similarly to the PMo CRN. From the bottom to the top, nuclearities are classified according to the number of Mo atoms. On the left side, species that only contain molybdenum are positioned, and the molecules with the same number of Mo, are separated according to the O/Mo ratio. In the center of the CRN, species that contain both molybdenum and arsenic are positioned. Last, on the right side, the two species that contain two arsenic atoms in their structure are placed.

In general terms, the reactions that form the more complex nuclearities are exergonic, except for the reaction that forms the $\{AsMo_9O_{31}\}$. This could explain why this species does not appear as in the speciation diagrams as much as the $\{AsMo_9O_{34}\}$. On the other hand, the reactions that form smaller nuclearities are more inhomogeneous. Among them, the reaction that forms the $\{Mo_3O_{11}\}$ species shows a high negative free energy. This species then reacts with another trimer structure to form the $\{Mo_6O_{21}\}$ species, which is the precursor of the $\{AsMo_9O_{34}\}$ nuclearity.

Concerning the formation of the $\{As_2Mo_5\}$, the cascade of reactions that form this species start with the formation of the $\{AsMo_3\}$ species,

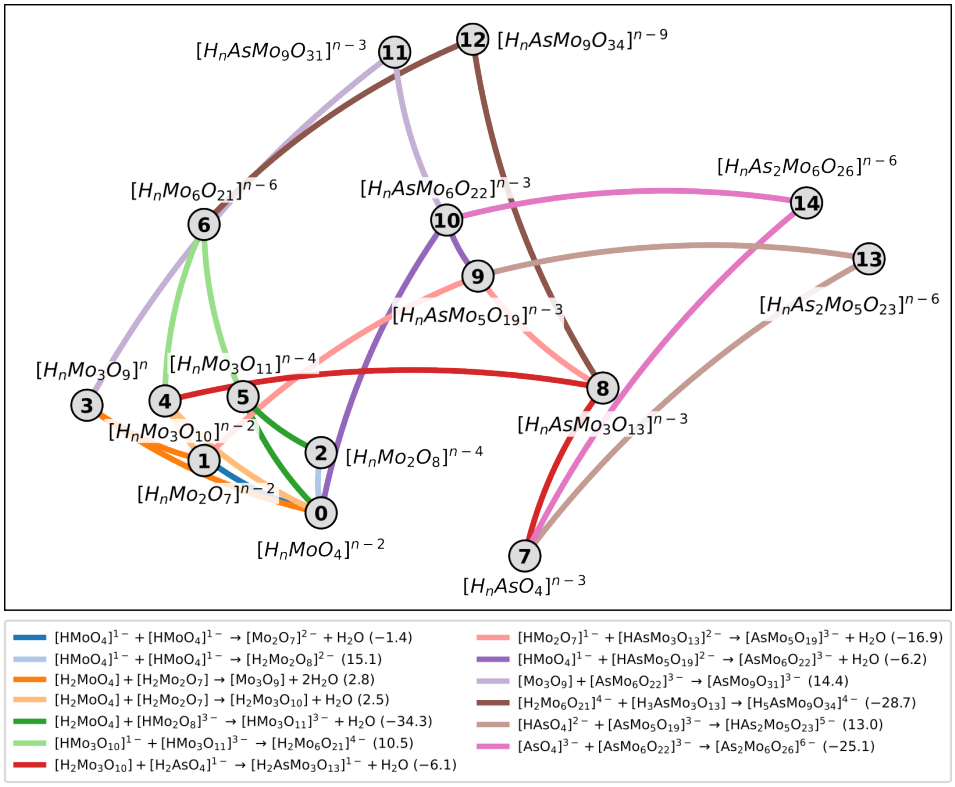

Figure 5.12: Chemical Reaction Network of the most frequent reactions in the selected 60k speciation models.

which then reacts with a dimer to form the $\{AsMo_5\}$ species. This molecule is a precursor to both $\{As_2Mo_5\}$ and $\{As_2Mo_6\}$ species. To form the Strandberg anion, it must first associate with another arsenate anion in a non-favorable reaction, that can take place for the accumulated energy from the previous reactions. For the $\{As_2Mo_6\}$, the precursor reacts with monomeric molybdenum anion, turning a 5-molybdenum ring to a 6-member one to later react with a new $\{AsO_4\}$ anion. These two re-

actions together are relatively favorable, and could explain the abundant presence of this molecule in the speciation of the AsMo system at low metal/heteroatom ratios.

5.4 Conclusions

In this chapter we employed the universal scaling algorithm together with the statistical pipeline described in *Chapter 3* to successfully predict the aqueous speciation for the arsenomolybdate system at various [Mo]/[As] concentration ratios.

We have validated the universal scaling methodology by comparing the experimental speciation diagrams with the predicted ones at different concentration ratios. We have reproduced the inverse behaviour in the speciation diagrams of the $\{As_2Mo_6\}$ and $\{AsMo_9\}$ species at distinct concentration ratios.

Moreover, we have unveiled the speciation phase diagrams for the arsenomolybdate, giving a general overview of the speciation from either the point of view of the molybdenum-based species or the phosphorus ones.

We have used the statistical pipeline to reduce the complexity of the Chemical Reaction Network of the arsenomolybdate system (96 reactions). The provided CRN contains the most frequent reaction forming each nuclearity highlighting the importance of the intermediate species to form the more complex clusters.

Chapter 6

Conclusions

In this thesis we have broadened the POMSimulator methodology to extend the applicability from isopolyoxometalates to heteropolyoxometalate systems, and increased the robustness by implementing statistical techniques to handle datasets of high dimensionality. We have defined and included the necessary equations to automatically generate chemical reaction networks for heteropolyoxometalates, dealing with the mass balance equations for both the metal and the heteroatom. To apply POMSimulator to large and complex systems we have improved the overall performance of the code accomplishing a 20-fold speed-up in the solution of the systems of non-linear equations associated with speciation models, reducing one of the main bottlenecks of the software.

Embracing the FAIR (Findability, Accessibility, Interoperability and Reusability) principles, we have released a first open-source version of the POMSimulator software, as well as published the DFT-characterized molecular sets in the ioChem-BD repository.

One of the limitations of POMSimulator was the need of experimental formation constants to perform a linear scaling of the DFT-calculated ones. To tackle this issue we have developed a universal scaling methodology, enabling the study of systems for which experimental formation constants are not reported. The slope parameter of the linear regression has been proven to be constant (m=0.29) confirming previous hypotheses. We have demonstrated that this value remains universal across different DFT methods and for multiple metal systems including IPAs and HPAs. Moreover, we have implemented a Multi-Linear Regression model to predict the system-dependent intercept parameter using only POMSimulator-derived data, reducing the dependence on quantitative experimental information.

We have defined a new protocol to clusterize large amounts of speciation models based on their speciation diagrams, to increase the robustness of POMSimulator. With this new approach, we can now obtain the error associated with our prediction, gaining the capability to assess the reliability of the speciation. Consequently, we have adopted a data-driven approach aligned with the increasing complexity of the generated datasets.

Combining the universal scaling with the statistical analysis of speciation models we have successfully predicted the aqueous speciation for two different HPA systems: the PMo and the AsMo.

For the phosphomolybdate system we have reported the formation of the Keggin anion at acidic pH as well as the Keggin lacunary and Strandberg anions at higher pH values, agreeing with the experimental speciation diagram. We have simulated the first speciation phase diagram for an HPA system, representing the metal/heteroatom concentration ratios versus the pH. Phase diagrams have proven to be an exceptional tool to decypher the intricate speciation of polyoxometalates at different conditions, in this case,

the amount of phosphorus and molybdenum present in solution.

Concerning the arsenomolybdate speciation, we have predicted the formation of the different protonation states of the $\{As_2Mo_6\}$ and $\{AsMo_9\}$ nuclearities at several concentration ratios. We have also reported the speciation phase diagram for the AsMo system to compare the information provided by this representation with the one in the speciation diagrams. Combining the general overview from phase diagrams with the detailed perspective from speciation diagrams, expands the understanding on the chemical behaviour of the system.

Moreover, this data-driven perspective paves the way for a more thorough analysis of the Chemical Reaction Network characteristic of polyoxometalate systems, providing better mechanistic insights.

To sum up, this work sets up the foundations for further developments of the POMSimulator methodology to include features such as oxidation and reduction reactions, non-aqueous solvents and the development of Machine Learning models.

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builes Casasnovas

Appendix A

Computational Details

The molecular geometries of all oxo-clusters were fully optimized employing the ADF software package (SCM ADF version 2019.1) [111], using the PBE functional [112, 113], with the relativistic corrections related to the scalar-relativistic zero-order regular approximation (ZORA) [114, 115], using a TZP basis set level. Solvation effects were introduced by means of the continuous solvent model COSMO with Klamt radii for water [116]. Stationary points were characterized with analytic frequency calculations. All Gibbs free energies were computed at 298.15 K and 1 atm, using the ideal gas-rigid rotor-harmonic oscillator (IGRRHO) model. QTAIM analysis was performed at the same level of theory. Single point energy calculations using the BP86 [112, 113, 117], B3LYP [118, 119] and M06L [120] functionals were also computed from the PBE optimized geometries. PBE thermochemical parameters were also employed to compute Gibbs free energies.

Protonation site selection

Polyoxometalates chemistry is highly dependent on the pH, and for this reason it is vital for the POMSimulator framework to properly define the acid-base chemistry of these molecules. In this sense, acid-base reactions are highly affected by the position in which protons are located in the molecule. An exhaustive study about the protonation sites of polyoxometalates would be the path to go, but it would also be very time-demanding. To tackle this issue, we decided to simplify the problem, and use a faster approach. We generated the molecular electrostatic potential (MEP), to discern which were the most basic sites, more prone to be protonated. The MEP of two polyoxometalates from the PMo system are depicted in Figure A.1. From these MEPs, we can then select the protonation sites.

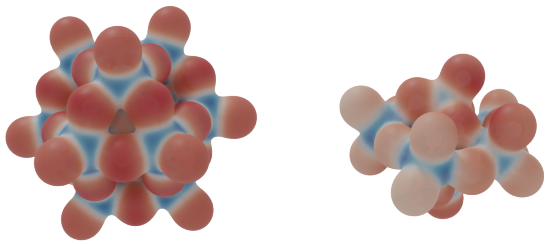

Figure A.1: Molecular electrostatic potentials for the Keggin anion (left) and the Strandberg anion (right). Red zones represent the most negative points in the electrostatic potential, and thus the most basic sites of the molecule. Blue zones represent the most positive points, and thus the most acidic sites.

Bibliography

(1) Berzelius, J. J. *Annalen der Physik* **1826**, *82*, 369–392.

(2) Keggin, J. F. *Nature* **1933**, *131*, 908–909.

(3) Anderson, J. S. *Nature* **1937**, *140*, 850–850.

(4) Evans, H. T. *Journal of the American Chemical Society* **1948**, *70*, 1291–1292.

(5) Dawson, B. *Acta Crystallographica* **1953**, *6*, 113–126.

(6) Lindqvist, I. *Arkiv for Kemi* **1953**, *5*, 247–250.

(7) Strandberg, R.; Niinistö, L.; Møller, J.; Schroll, G.; Leander, K.; Swahn, C.-G. *Acta Chemica Scandinavica* **1973**, *27*, 1004–1018.

(8) Pope, M. T., *Heteropoly and Isopoly Oxometalates*, Series Title: Inorganic Chemistry Concepts; Springer Berlin Heidelberg: Berlin, Heidelberg, 1983; Vol. 8.

(9) Müller, A.; Krickemeyer, E.; Meyer, J.; Bögge, H.; Peters, F.; Plass, W.; Diemann, E.; Dillinger, S.; Nonnenbruch, F.; Randerath, M.; Menke, C. *Angewandte Chemie International Edition* **1995**, *34*, 2122–2124.

(10) Caruso, F.; Kurth, D. G.; Volkmer, D.; Koop, M. J.; Müller, A. *Langmuir* **1998**, *14*, 3462–3465.

(11) Web Of Science search `https://www.webofscience.com/wos/alldb/summary/c20b4b00-63cb-4156-b76d-748f6a6b12ad-f65299a7/relevance/1` (accessed 05/21/2024).

(12) Hervé, G.; Tézé, A.; Contant, R. In *Polyoxometalate Molecular Science*, Borrás-Almenar, J. J., Coronado, E., Müller, A., Pope, M., Eds.; Springer Netherlands: Dordrecht, 2003, pp 33–54.

(13) Miras, H.; Long, D.-L.; Cronin, L. In *Advances in Inorganic Chemistry*; Elsevier: 2017; Vol. 69, pp 1–28.

(14) Gumerova, N. I.; Rompel, A. *Chemical Society Reviews* **2020**, *49*, Publisher: Royal Society of Chemistry, 7568–7601.

(15) Gumerova, N. I.; Rompel, A. *Science Advances* **2023**, *9*, eadi0814.

(16) Bijelic, A.; Aureliano, M.; Rompel, A. *Angewandte Chemie International Edition* **2019**, *58*, 2980–2999.

(17) Song, N.; Lu, M.; Liu, J.; Lin, M.; Shangguan, P.; Wang, J.; Shi, B.; Zhao, J. *Angewandte Chemie International Edition* **2024**, e202319700.

(18) Khoshkhan, Z.; Mirzaei, M.; Amiri, A.; Lotfian, N.; Mague, J. T. *Inorganic Chemistry* **2024**, DOI: `10.1021/acs.inorgchem.3c02130`.

(19) Dan, K.; Fujinami, K.; Sumitomo, H.; Ogiwara, Y.; Suhara, S.; Konno, Y.; Sawada, M.; Soga, Y.; Takada, A.; Takanashi, K.; Watanabe, K.; Shinozuka, T. *Applied Sciences* **2020**, *10*, 8246.

(20) Lan, J.; Wang, Y.; Huang, B.; Xiao, Z.; Wu, P. *Nanoscale Advances* **2021**, *3*, 4646–4658.

(21) Solé-Daura, A.; Benseghir, Y.; Ha-Thi, M.-H.; Fontecave, M.; Mialane, P.; Dolbecq, A.; Mellot-Draznieks, C. *ACS Catalysis* **2022**, *12*, 9244–9255.

(22) Gusmão, F. M. B.; Mladenović, D.; Radinović, K.; Santos, D. M. F.; Šljukić, B. *Energies* **2022**, *15*, 9021.

(23) Zhang, Y.; Li, Y.; Guo, H.; Guo, Y.; Song, R. *Materials Chemistry Frontiers* **2024**, *8*, 732–768.

(24) Chen, J.-J.; Vilà-Nadal, L.; Solé-Daura, A.; Chisholm, G.; Minato, T.; Busche, C.; Zhao, T.; Kandasamy, B.; Ganin, A. Y.; Smith, R. M.; Colliard, I.; Carbó, J. J.; Poblet, J. M.; Nyman, M.; Cronin, L. *Journal of the American Chemical Society* **2022**, *144*, 8951–8960.

(25) Molina, P. I.; Sures, D. J.; Miró, P.; Zakharov, L. N.; Nyman, M. *Dalton Transactions* **2015**, *44*, 15813–15822.

(26) Kikkawa, S.; Tsukada, M.; Shibata, K.; Fujiki, Y.; Shibusawa, K.; Hirayama, J.; Nakatani, N.; Yamamoto, T.; Yamazoe, S. *Symmetry* **2021**, *13*, 1267.

(27) Albert, J.; Mehler, J.; Tucher, J.; Kastner, K.; Streb, C. *ChemistrySelect* **2016**, *1*, 2889–2894.

(28) Pettersson, L.; Andersson, I.; Óhman, L.-O. *Inorganic Chemistry* **1986**, 4726–4733.

(29) Pettersson, L.; Carlsson, B.; Rundqvist, S.; Andresen, A. F.; Fischer, P. *Acta Chemica Scandinavica* **1975**, *29a*, 677–689.

(30) Attique, S.; Batool, M.; Yaqub, M.; Gregory, D. H.; Wilson, C.; Goerke, O.; Shah, A. T. *Materials Chemistry and Physics* **2020**, *246*, 122781.

(31) Han, Z.-G.; Chang, X.-Q.; Yan, J.-S.; Gong, K.-N.; Zhao, C.; Zhai, X.-L. *Inorganic Chemistry* **2014**, *53*, 670–672.

(32) Öhman, L.-O.; Nordin, A.; Rømming, C.; Røst, E.; Khan, A. Z.-Q.; Sandström, J.; Krogsgaard-Larsen, P. *Acta Chemica Scandinavica* **1992**, *46*, 515–520.

(33) Coronel, N. C.; Da Silva, M. J. *Journal of Cluster Science* **2018**, *29*, 195–205.

(34) Ni, L.; Spingler, B.; Weyeneth, S.; Patzke, G. R. *European Journal of Inorganic Chemistry* **2013**, *2013*, 1681–1692.

(35) Long, D.-L.; Burkholder, E.; Cronin, L. *Chemical Society Reviews* **2007**, *36*, 105–121.

(36) Wilson, E. F.; Miras, H. N.; Rosnes, M. H.; Cronin, L. *Angewandte Chemie International Edition* **2011**, *50*, 3720–3724.

(37) Winter, R. S.; Cameron, J. M.; Cronin, L. *Journal of the American Chemical Society* **2014**, *136*, 12753–12761.

(38) Corella-Ochoa, M. N.; Miras, H. N.; Long, D.-L.; Cronin, L. *Chemistry A European Journal* **2012**, *18*, 13743–13754.

(39) Yao, S.; Falaise, C.; Leclerc, N.; Roch-Marchal, C.; Haouas, M.; Cadot, E. *Inorganic Chemistry* **2022**, *61*, Publisher: American Chemical Society, 4193–4203.

(40) Pascual-Borràs, M.; López, X.; Rodríguez-Fortea, A.; Errington, R. J.; Poblet, J. M. *Chemical Science* **2014**, *5*, 2031–2042.

(41) Fournier, M.; Thouvenot, R.; Rocchiccioli-Deltcheff, C. *Journal of the Chemical Society, Faraday Transactions* **1991**, *87*, 349–356.

(42) Grama, L.; Boda, F.; Gaz Florea, A.; Curticăpean, A.; Muntean, D.-L. *Acta Medica Marisiensis* **2014**, *60*, 84–88.

(43) Teague, C. M.; Li, X.; Biggin, M. E.; Lee, L.; Kim, J.; Gewirth, A. A. *Journal of Physical Chemistry B* **2004**, *108*, 1974–1985.

(44) Kuepper, K.; Derks, C.; Taubitz, C.; Prinz, M.; Joly, L.; Kappler, J.-P.; Postnikov, A.; Yang, W.; Kuznetsova, T. V.; Wiedwald, U.; Ziemann, P.; Neumann, M. *Dalton Transactions* **2013**, *42*, 7924.

(45) Hou, Y.; Zakharov, L. N.; Nyman, M. *Journal of the American Chemical Society* **2013**, *135*, 16651–16657.

(46) Li, M.; Zheng, Z.; Yin, P. *Journal of Coordination Chemistry* **2020**, *73*, 2365–2372.

(47) Marques, M. P. M.; Gianolio, D.; Ramos, S.; Batista De Carvalho, L. A. E.; Aureliano, M. *Inorganic Chemistry* **2017**, *56*, 10893–10903.

(48) Falbo, E.; Rankine, C.; Penfold, T. *Chemical Physics Letters* **2021**, *780*, 138893.

(49) Matsuyama, T.; Kikkawa, S.; Kawamura, N.; Higashi, K.; Nakatani, N.; Kato, K.; Yamazoe, S. *Journal of Physical Chemistry C* **2024**, *128*, 2953–2958.

(50) Yao, S.; Falaise, C.; Khlifi, S.; Leclerc, N.; Haouas, M.; Landy, D.; Cadot, E. *Inorganic Chemistry* **2021**, *60*, 7433–7441.

(51) Taketa, H.; Katsuki, S.; Eguchi, K.; Seiyama, T.; Yamazoe, N. *Journal of Physical Chemistry* **1986**, *90*, 2959–2962.

(52) Web Of Science search `https://www.webofscience.com/wos/alldb/summary/59e09631-b50f-48f2-a9a0-fa4d8741bd2e-ea5c6772/relevance/1` (accessed 05/21/2024).

(53) Ritschl, F.; Haberlandt, H. *Journal of Molecular Structure: THEOCHEM* **1988**, *180*, 45–63.

(54) Kempf, J. Y.; Rohmer, M. M.; Poblet, J. M.; Bo, C.; Benard, M. *Journal of the American Chemical Society* **1992**, *114*, 1136–1146.

(55) Poblet, J. M.; López, X.; Bo, C. *Chemical Society Reviews* **2003**, *32*, 297–308.

(56) Kibler, A. J.; Tsang, N.; Winslow, M.; Argent, S. P.; Lam, H. W.; Robinson, D.; Newton, G. N. *Inorganic Chemistry* **2023**, *62*, 3585–3591.

(57) Kondinski, A. *Nanoscale* **2021**, *13*, 13574–13592.

(58) Chi, M.; Zeng, Y.; Lang, Z.-L.; Li, H.; Xin, X.; Dong, Y.; Fu, F.; Yang, G.-Y.; Lv, H. *ACS Catalysis* **2024**, *14*, 5006–5015.

(59) Salazar Marcano, D. E.; Savić, N. D.; Abdelhameed, S. A. M.; De Azambuja, F.; Parac-Vogt, T. N. *JACS Au* **2023**, *3*, 978–990.

(60) Malcolm, D.; Vilà-Nadal, L. *ACS Organic and Inorganic Au* **2023**, *3*, 274–282.

(61) Stergiou, A. D.; Symes, M. D. *Catalysis Today* **2022**, *384-386*, 146–155.

(62) Thompson, J. A.; Vilà-Nadal, L. *Dalton Transactions* **2024**, *53*, 564–571.

(63) Bridgeman, A. J. *Chemistry A European Journal* **2004**, *10*, 2935–2941.

(64) Bridgeman, A. J. *Chemistry A European Journal* **2006**, *12*, 2094–2102.

(65) Centellas, M. S.; Piot, M.; Salles, R.; Proust, A.; Tortech, L.; Brouri, D.; Hupin, S.; Abécassis, B.; Landy, D.; Bo, C.; Izzet, G. *Chemical Science* **2020**, *11*, 11072–11080.

(66) Falbo, E.; Penfold, T. J. *Journal of Physical Chemistry C* **2020**, *124*, 15045–15056.

(67) López, X.; Nieto-Draghi, C.; Bo, C.; Avalos, J. B.; Poblet, J. M. *Journal of Physical Chemistry A* **2005**, *109*, 1216–1222.

(68) Lowe, M. P.; Lockhart, J. C.; Forsyth, G. A.; Clegg, W.; Fraser, K. A. *Journal of the Chemical Society Dalton Transactions* **1995**, 145.

(69) Li, Z.; Huang, Y.; Li, H.; Zhang, F.; Ren, Y.; Shi, W.; Liu, Q.; Wang, X. *Journal of the American Chemical Society* **2024**, *146*, 450–459.

(70) Courcot, B.; Bridgeman, A. J. *Journal of Computational Chemistry* **2011**, *32*, 3143–3153.

(71) Vilà-Nadal, L.; Mitchell, S. G.; Rodríguez-Fortea, A.; Miras, H. N.; Cronin, L.; Poblet, J. M. *Physical Chemistry Chemical Physics* **2011**, *13*, 20136.

(72) Petrus, E.; Segado, M.; Bo, C. *Chemical Science* **2020**, *11*, 8448–8456.

(73) Petrus, E.; Bo, C. *Journal of Physical Chemistry A* **2021**, *125*, 5212–5219.

(74) Petrus, E.; Segado-Centellas, M.; Bo, C. *Inorganic Chemistry* **2022**, *61*, Publisher: American Chemical Society, 13708–13718.

(75) Petrus, E.; Garay-Ruiz, D.; Reiher, M.; Bo, C. *Journal of the American Chemical Society* **2023**, *145*, 18920–18930.

(76) Hohenberg, P.; Kohn, W. *Physical Reviews* **1964**, *136*, B864–B871.

(77) Kohn, W.; Sham, L. J. *Physical Reviews* **1965**, *140*, A1133–A1138.

(78) Perdew, J. P. In *AIP Conference Proceedings*; ISSN: 0094243X, AIP: Antwerp (Belgium), 2001; Vol. 577, pp 1–20.

(79) Bader, R. F. W. *Accounts of Chemical Research* **1985**, *18*, 9–15.

(80) Leonard Euler *Commentarii Academiae Scientiarum Imperialis Petropolitanae* **1736**, *8*, 128–140.

(81) Morán-González, L.; Betten, J. E.; Kneiding, H.; Balcells, D. AABBA: Atom–Atom Bond–Bond Bond–Atom Graph Kernel for Machine Learning on Molecules and Materials, 2023.

(82) Garay-Ruiz, D.; Bo, C. *Journal of Cheminformatics* **2022**, *14*, 29.

(83) Pablo-García, S.; Pérez-Soto, R.; Sabadell-Rendón, A.; Garay-Ruiz, D.; Nosylevskyi, V.; López, N. *Digital Discovery* **2024**, 10.1039.D4DD00087K.

(84) Wen, M.; Spotte-Smith, E. W. C.; Blau, S. M.; McDermott, M. J.; Krishnapriyan, A. S.; Persson, K. A. *Nature Computational Science* **2023**, *3*, 12–24.

(85) Unsleber, J. P.; Reiher, M. *Annual Review of Physical Chemistry* **2020**, *71*, 121–142.

(86) Simm, G. N.; Reiher, M. *Journal of Chemical Theory and Computation* **2017**, *13*, 6108–6119.

(87) Garay-Ruiz, D.; Álvarez-Moreno, M.; Bo, C.; Martínez-Núñez, E. *ACS Physical Chemistry Au* **2022**, *2*, 225–236.

(88) Simm, G. N.; Reiher, M. *Journal of Chemical Theory and Computation* **2017**, *13*, 6108–6119.

(89) Unsleber, J. P.; Grimmel, S. A.; Reiher, M. *Journal of Chemical Theory and Computation* **2022**, *18*, 5393–5409.

(90) Petrus, E.; Builds, J.; Garay-Ruiz, D.; Segado-Centellas, M.; Bo, C. petrusen/pomsimulator: Release 1.0.0, 2024.

(91) Miras, H. N.; Cooper, G. J. T.; Long, D.-L.; Bögge, H.; Müller, A.; Streb, C.; Cronin, L. *Science* **2010**, *327*, 72–74.

(92) Pettersson, L.; Carlsson, B.; Rundqvist, S.; Andresen, A. F.; Fischer, P. *Acta Chemica Scandinavica* **1975**, *29a*, 677–689.

(93) Raabe, J.-C.; Jameel, F.; Stein, M.; Albert, J.; Poller, M. J. *Dalton Transactions* **2024**, *53*, 454–466.

(94) Zhang, X.; Luo, X.; Duan, Y.; Huang, Y.; Zhang, N.; Zhao, L.; Wu, J. *Journal of Molecular Structure* **2017**, *1141*, 245–251.

(95) Zhang, M.; Lv, J.; Yu, K.; Wang, K.; Meng, F.; Zhou, B. *Inorganic Chemistry Communications* **2018**, *97*, 74–78.

(96) Himeno, S.; Hashimoto, M.; Ueda, T. *Inorganica Chimica Acta* **1999**, *284*, 237–245.

(97) Rozantsev, G. M.; Sazonova, O. I. *Russian Journal of Coordination Chemistry* **2005**, *31*, 552–558.

(98) Elvingson, K.; González Baró, A.; Pettersson, L. *Inorganic Chemistry* **1996**, *35*, 3388–3393.

(99) Khalaji-Verjani, M.; Masteri-Farahani, M. *ACS Applied Energy Materials* **2024**, *7*, 6612–6620.

(100) Martins, I. C. B.; Al-Sabbagh, D.; Bentrup, U.; Marquardt, J.; Schmid, T.; Scoppola, E.; Kraus, W.; Stawski, T. M.; Guilherme Buzanich, A.; Yusenko, K. V.; Weidner, S.; Emmerling, F. *Chemistry A European Journal* **2022**, *28*, e202200079.

(101) Raabe, J.-C.; Esser, T.; Jameel, F.; Stein, M.; Albert, J.; Poller, M. J. *Inorganic Chemistry Frontiers* **2023**, *10*, 4854–4868.

(102) Pettersson, L.; Andersson, I.; Óhman, L.-O. *Inorganic Chemistry* **1986**, 4726–4733.

(103) Sun, F.; Yang, L.; Yue, C.; Liu, Y.; Bao, W.; Tuo, Y.; Feng, X.; Lu, Y. *International Journal of Hydrogen Energy* **2022**, *47*, 25571–25582.

(104) Li, F.-R.; Ji, T.; Chen, W.-L. *Tungsten* **2022**, *4*, 99–108.

(105) Wang, S.-H.; Jansen, S. A. *MRS Proceedings* **1994**, *368*, 229.

(106) López, X.; Poblet, J. M. *Inorganic Chemistry* **2004**, *43*, 6863–6865.

(107) Builds, J.; Garay-Ruiz, D.; Petrus, E. ioChem-BD Data Collection: Phosphomolybdate molecular set `https://doi.org/10.19061/iochem-bd-1-323` (accessed 10/01/2024).

(108) Álvarez-Moreno, M.; De Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. *Journal of Chemical Information and Modeling* **2015**, *55*, 95–103.

(109) Bo, C.; Maseras, F.; López, N. *Nature Catalysis* **2018**, *1*, 809–810.

(110) Builds, J.; Garay-Ruiz, D.; Petrus, E. ioChem-BD Data Collection: Arsenomolybdate molecular set `https://doi.org/10.19061/iochem-bd-1-346` (accessed 10/01/2024).

(111) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J.; Snijders, J. G.; Ziegler, T. *Journal of Computational Chemistry* **2001**, *22*, 931–967.

(112) Perdew, J. P. *Physical Review B* **1986**, *33*, 8822–8824.

(113) Perdew, J. P. *Physical Review B* **1986**, *34*, 7406.

(114) Van Lenthe, E.; Baerends, E. J.; Snijders, J. G. *The Journal of Chemical Physics* **1993**, *99*, 4597–4610.

(115) Van Lenthe, E.; Baerends, E. J. *Journal of Computational Chemistry* **2003**, *24*, 1142–1156.

(116) Klamt, A. *The Journal of Chemical Physics* **1995**, *99*, 2224–2235.

(117) Becke, A. D. *Physical Review A* **1988**, *38*, 3098–3100.

(118) Lee, C.; Yang, W.; Parr, R. G. *Physical Review B* **1988**, *37*, 785–789.

(119) Becke, A. D. *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

(120) Zhao, Y.; Truhlar, D. G. *The Journal of Chemical Physics* **2006**, *125*, 194101.

UNIVERSITAT ROVIRA I VIRGILI
Enhancing POMSimulator Applications in Heteropolyoxometalates: A Statistical and Data-Driven Approach
Jordi Builds Casasnovas