Representations and descriptors unifying the study of molecular and bulk systems

PERSPECTIVE

Representations and descriptors unifying the study of molecular and bulk systems

K. Rossi[1]* | **J. Cumby**[2]*

[1]Laboratory of Computational Science and Modeling, Institute des Materiaux, Ecole Polytechnique Federale de Lausanne, CH-1015 Lausanne, Switzerland

[2]School of Chemistry, University of Edinburgh, Edinburgh, EH9 3FD, UK

Correspondence
James Cumby, School of Chemistry, University of Edinburgh, Edinburgh, EH9 3FD, UK
Email: james.cumby@ed.ac.uk

Funding information
KR acknowledges funding from EPFL and Solvay.

Establishing a unified framework for describing the structures of molecular and periodic systems is a long-standing challenge in physics, chemistry, and material science. With the rise of machine learning methods in these fields, there is a growing need for such a method. This perspective aims to discuss the development and use of three promising approaches - topological, atom-density, and symmetry-based - for the prediction and rationalisation of physical, chemical and mechanical properties of atomistic systems across different scales and compositions.

KEYWORDS
Descriptors, Machine Learning, Connectivity, Symmetry-Distortions, Atom-Density, Data-Driven

1 | INTRODUCTION

The attempt to infer a mapping between the atomic structure of a system and its chemical and physical properties - and thus also the associated methods to describe chemical structures - have a long history, with examples dating back to the mid 19th century. [1] Quantum mechanics provides a universal framework for computing the electronic structure and properties of materials and molecules alike, however, it can be computationally prohibitive depending on the size of the system and the number of approximations involved. Machine learning (in some cases combined with quantum mechanical methods) promises to provide an alternative framework for property predictions in atomistic systems. The downside of machine learning approaches is that they are often restricted to limited classes of materials or physical properties, lacking the desired universality. The development of transferable ways to describe materials for data-driven techniques are therefore essential for the development of efficient and accurate routines for rationalising the behaviour of molecular and crystalline systems, and the design of improved technological devices and materials.

* Equally contributing authors.

While a list of Cartesian atomic coordinates serves to completely and uniquely define a chemical structure, permuting atom labels or applying an arbitrary rotation would render a completely different representation of the system, although the underlying structure has not changed. As such, there is a need for structural *descriptors* or *representations* (we shall use these terms interchangeably) for both molecules and extended solids, which summarise structural information in a way suitable for direct comparison between systems. A general definition of a structural descriptor is one that is invariant to rotation, translation, reflection (unless chiral isomers shall be distinguished) or atomic permutation operations. Uniqueness and invertibility (the ability to reproduce the original structure from a descriptor) are also often invoked as key for a good descriptor, but these criteria are not essential for the success of a representation framework for application in data-driven approaches.

In this perspective we discuss some of the recent progress in the broad area of research concerning the development of approaches unifying the study of atomistic systems of different scales and compositions, ranging from small molecules to extended solids. Because of the subject's rapid and continuous development we will not attempt to cover all the achievements in the field, but will try to offer the reader a bird's-eye view which highlights some of the key successes, and attempt to identifying open routes for further development. We focus on descriptors based on structural properties, on the premise that the atomic structure is inherently linked to the resulting physical properties observed. The manuscript is arranged as follows: first topological descriptors and graphs arising from these are presented, secondly the development and use of atomic density descriptors are analysed, and finally a discussion on distortion-based descriptors is brought forward.

2 | TOPOLOGICAL DESCRIPTORS

Historically, molecular graphs were the first descriptors introduced in the study of chemical systems. According to the IUPAC definition this is a labelled graph where each atom in the molecule is represented by a node and each chemical bond corresponds to an edge connecting two nodes,[2] see also Figure 1 for a representation of how graphs are drawn from a chemical structure. Molecular graphs encode only topological properties, *i.e.* a notion of pairwise bonding relations, and disregard distance or angular properties, *e.g.* the notion of whether a *cis-* or *trans-* configuration is observed in the molecule of interest. Graphs naturally result in translation, rotation, and permutation invariant representations, but at a loss of uniqueness, *i.e.*, different isomers of the same molecule are mapped on to the same graph. To obviate this problem graph edges and nodes may be labelled with features such as bond lengths or absolute configuration (R/S nomenclature). Interestingly, molecular graphs were most successfully introduced by Arthur Cayley with earlier attempts in the mid of 18th century, even in the absence of a full notion of chemical bonding. [1]

Fast forward to the 21st century and molecular graphs are commonly used in supervised machine learning algorithms for the high-throughput prediction of several physico-chemical properties, with interest ranging from applications in biomedicine to solid-state device manufacturing. The framework of graph-based models for molecules is in fact naturally suited for carrying out predictions in message-passing neural network schemes, which are a subset of the growing range of geometric deep learning methods. These methods extend the approach of neural networks into operating on non-Euclidean data (such as graphs) directly, rather than first transforming it to a vectorial representation. Examples of the strong predictive power of molecular graphs include for physical properties such as dipole moment and heat capacity as reported in the QM9 data set[3, 4] and high-throughput polymer screening. [5]

Formal grammars for representing molecules based on their graph - as in SMILES [6] or InChI [7] - provide a natural extension, and are particularly useful given the maturity of natural language processing techniques. Such grammars (of which there are many) are based on tree-like traversal of molecules, and therefore reconstructing the

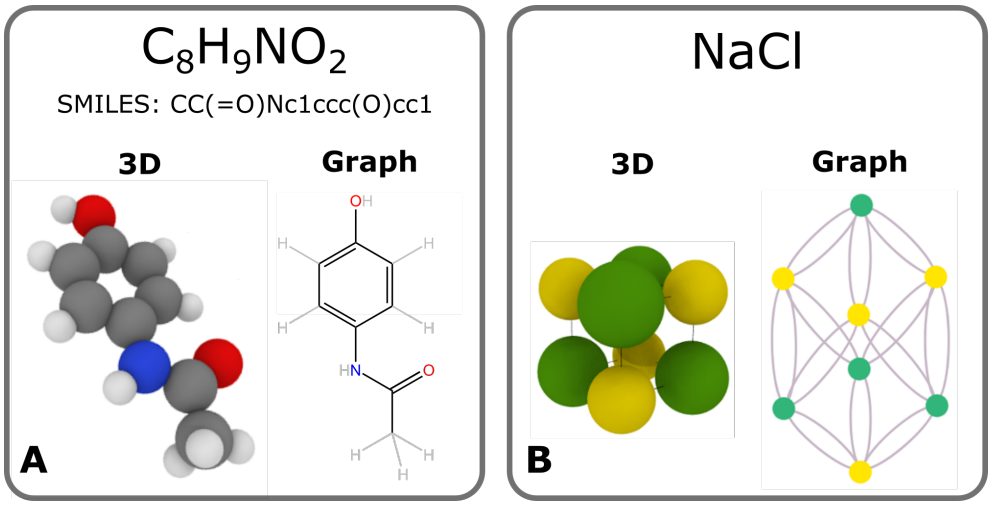

FIGURE 1 Graphical example of how 3D structures for (A) molecular and (B) periodic systems can be translated into corresponding graphs. SMILES notation is also reported in (A) for the corresponding paracetamol molecule. Only the asymmetric unit of the periodic structure is shown in (B).

original molecular topology requires persistent knowledge of the tree for which recurrent algorithms are essential. Notable examples of their use are found for the prediction of molecular properties ranging from atomization energies to biological activities[8] as well as *de novo* drug design [9, 10] and retro-synthetic routes. [11, 12]

The rigorous definition of a coordination shell in bulk systems (such as through Voronoi tessellation) also allows for the development of periodic (crystalline) graphs. Parallel to the case of molecular ones, each atom in the asymmetric unit cell is a node and connections between atoms are tagged by edges. Multiple connections between two nodes are allowed to account for the periodicity intrinsic to crystalline systems (fig. 1B). Crystal graph based approaches - also possibly encoding system features by labelling nodes, edges and global states - have been successful in predicting formation energies, band gaps, and mechanical properties with great accuracy of several thousands of crystals with various compositions, either by using the resulting graph construction directly as input to a geometric neural network discussed above,[13] or by concatenating smaller sub-graphs representing local environments into a more traditional feature vector for use in a decision tree method. [14]

While graphs and connectivity grammar approaches have demonstrated their flexibility in the study of molecular systems and well-defined crystals, their application in finite-size inorganic systems, *e.g.* metallic nanoparticle and organometallic systems has been limited. The successes in encoding a linear relationship between a coordination measure and adsorption energies[15] and the advancement of site counting schemes in predicting catalytic properties [16, 17] anticipate the potential of such methods in heterogeneous catalysis. Similarly, the combination of topological descriptors with regression schemes have already found successful applications in rationalising the selectivity and activity of reactions catalysed by organometallic complexes. [18, 19] By the same token, we note that the use of graph set analysis for structural pattern recognition for both supervised and unsupervised tasks, strongly developed in the past literature,[20] is observing a renewed surge in interest. These advancements encompass, for example, the development of topological coordinates which are permutation invariant [21] as well as the discussion of novel graph matching schemes explicitly adapted to tackle problems related to simulations of materials. [22, 23]

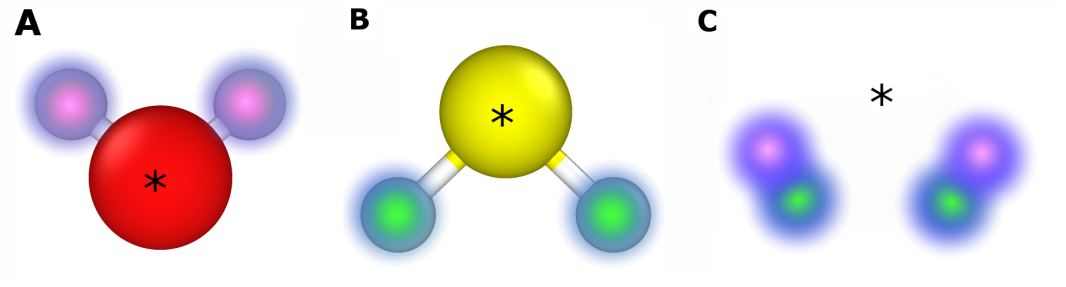

FIGURE 2 Pictorial representation of how two local environments for an oxygen (A) and a sulfur (B) in a H_2O and a H_2S molecule respectively can be compared by contrasting a carefully chosen rotation, translation and permutation invariant abstract representation (C). By focusing on the heavy atom (oxygen in A and sulfur in B, denoted by $*$) and generating an abstract representation of its surrounding which is rotationally, permutationally, and translationally invariant (shown by the coloured hue), the overlap between the $*$-centred representations and the similarity between the two can be evaluated.

3 | ATOM-DENSITY DESCRIPTORS

Many physical properties originate primarily due to short-range connectivity or bonding between nearest-neighbour atoms. As such, describing a system as a collection of local atomic environments is a sufficient condition for obtaining accurate predictions. Within this premise, atom-density representations of local environments have attracted a huge interest recently as a novel unifying approach in the study of atomic systems. Figure 2 shows an example of how density overlap kernels between atom-centred local environments allow quantitative analysis of the similarity between different molecules and atomic environments. Within this framework local environments are characterised by means of high dimensional feature spaces which accurately encode the relevant density distributions surrounding each atom in the system. Alternatively, comparison between materials in this higher dimensional representation can be achieved implicitly by employing the 'kernel trick', whereby the similarity in this other space can be quantified without performing the computationally difficult projection. Interested readers are referred to [24] for an overview in a chemical context. Such feature spaces can be equivalently built:[25, 26]

- from radial and spherical harmonic expansion of smoothed atomic densities via atom-centered Gaussians, i.e. the Smooth Overlap of Atomic Integrals representation, [27]
- through the arbitrary non-linear combination of 2- and 3-body descriptors, as in the case of Beheler-Parrinello symmetry functions, [28] Chebychev polynomial expansion of radial and angular distribution functions, [29] and scaled Gaussian [30] basis functions, or
- from functions of n-body kernels. [31, 25]

Other measures of densities and symmetries in the neighbourhood of atoms were previously presented in the literature, e.g. in the seminal work by Steinhardt and coworkers. [32] The widespread use of these descriptors in machine learning methods, however, stem from the need to incorporate a comprehensive representation in supervised learning schemes for forces and energies. Their successful application was then demonstrated across molecular and bulk systems of a variety of compositions. Their use is not limited to force and energy predictions, however, extending to tensorial properties,[33] pattern recognition for ice phases,[34] and atomic mobility of defects and grain boundaries. [35] The

generality of such a descriptor has also allowed it to be applied in unsupervised schemes to determine similarities among previously challenging-to-compare systems for the case of molecular systems in both crystalline and gas phases. [36] Atom-density descriptors used in semi-supervised multivariate analysis further allowed for the definition of a generalised convex hull construction. [37]

A drawback in the original formulation of atom-density representations lies in the combinatorial explosion in the feature space for the case of systems comprising several chemical species, which can result in excessive parameterisation for many problems. The use of alchemical representations and compositional descriptors proved helpful in alleviating this problem in practical applications such as unsupervised tasks [38] or energy prediction. [29] Furthermore, while atom-density descriptors have been remarkably accurate when applied to local property challenges, they cannot be of help in resolving non-local problems, e.g. predicting the energy of systems governed by long-range electrostatic interaction such as charged dimers. A recent advance in this area has shown how an equivalent long-range description, remapped as a feature vector defined locally and equivariant in O(3), could hold the solution. [39]

4 | DISTORTION DESCRIPTORS

Whilst the absolute configuration of atomic centres provides a robust input to machine learning algorithms, in many cases it is the displacements from a higher (or lower) symmetry configuration that better represent the underlying chemistry involved. For example, octahedral metal complexes such as $[Fe(bpy)_3]^{2+}$ (where bpy = 2,2'-bipyridine) exhibit two isomers with differing optical chirality and identical bond lengths, but can be related by a shared $D3h$ subgroup symmetry through the Bailar-twist mechanism. [40] Equally, materials undergoing a structural phase transition can exhibit two very different atomic arrangements, but often the transformation can be described by a single 'soft phonon' (such as the ferroelectric transition of $PbTiO_3$). The symmetry of local arrangements around atoms has frequently been invoked as a tool to characterise structural patterns and rationalise changes in the physico-chemical properties of materials or molecules.

Distorted coordination environments naturally lead to the question of how far from an ideal polyhedron is a bonding arrangement. One solution to this is continuous symmetry and shape measures, which provide a quantitative measure of how close a given polyhedron is to a reference (fig. 3A). [41, 42] Other related approaches include those based upon bond order parameters[32, 43] or other analytical expressions [44, 45] as well as template matching approaches based upon polyhedral fitting. [46, 47] A disadvantage of all of these schemes, however, is the requirement to chose a reference polyhedron, which is not necessarily trivial to determine *a priori*.

Another approach to quantify angle and bond length distortions simultaneously is based on fitting the minimum-volume ellipsoid which contains the entire polyhedron, resulting in three principal radii $r_1 \geq r_2 \geq r_3$ and a rotation matrix. [48]. This descriptor can be calculated for arbitrary coordination numbers, and has found application in understanding perovskite phase transitions and discovery of an off-centre d^5 effect in FeO_6 polyhedra from data-mining (figf. 3B). The associated shape parameter $S = \frac{r_3}{r_2} - \frac{r_2}{r_1}$ allows comparison of different coordination numbers on the same footing. [49].

A more complete description of local coordination (again requiring a reference object) is symmetry mode (or normal mode) analysis, where cooperative displacements of groups of atoms are quantified according to symmetry-derived motions, for instance the symmetric and asymmetric stretches in a linear triatomic molecule. Traditionally employed in the analysis of vibrational spectra, normal modes meet the invariance requirements of a materials descriptor, and faithfully reproduce the absolute atomic structure. While often used in descriptions of discrete molecules, they have found increasing use in extended solids, for instance in rationalising physical properties and phase transitions in

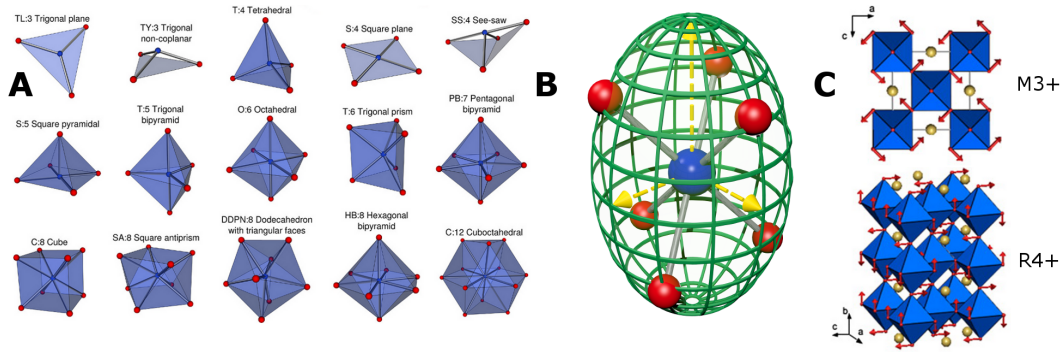

FIGURE 3 (A) Ideal coordination polyhedra commonly observed and used as reference shapes for bonding analysis (adapted with permission from [50]. Copyright 2017 American Chemical Society.); (B) representation of the ellipsoidal approach to quantifying coordination distortions; (C) Example symmetry adapted modes (irreps) describing distortions in ABO_3 perovskites (from [51]).

perovskite materials. [52]

Although commonly used to rationalise experimental data, symmetry analysis has so far seen limited application in data-driven approaches, such as predicting non-centrosymmetric perovskite-derived materials[53] or as a method to produce a non-redundant set of training structures for further machine learning. [54] The application of symmetry arguments to scientific problems has a long-proven and reliable track record, however, for instance in the broad applicability of molecular orbital (MO) theory. As such, it could enhance (and in many cases simplify) chemical machine learning problems. One reason for its limited acceptance so far may be the complication that symmetry is absolute (*i.e. a mirror plane either exists or doesn't*) so gradual atomic changes in molecules appear discontinuous. Similarly, not all structures can be described relative to a single 'parent' symmetry, therefore limiting comparisons between different structural classes. The development of a symmetry-based approach which is more gradual with coordination changes is a key problem to be addressed which, if solved, would increase systematic and robust machine learning between different chemical sub-disciplines (for instance small molecules and extended solids).

5 | OUTLOOK

Determining universal representations and descriptors for the study of finite-size and periodic systems is a high-reward challenge in physics, chemistry, and materials science. These empower general and flexible machine learning models for materials, whilst also bridging across sub-disciplines, for example linking small molecule properties with adsorption behaviours in porous solids. The three classes of methods discussed in this perspective (topology-based, atom-density-based, and distortion-based) demonstrate the leaps being made in this area of research.

All approaches mentioned have limitations. One general aspect to point out is that all of these descriptors are reliant on accurate structural information, neglecting the uncertainty inherent to physical measurements. An interesting open question is whether this uncertainty could be included in future descriptors to influence the resulting properties, for instance whether such uncertainty could lead to a statistically meaningful ensemble of predictions. Similarly, the complementary combination of different (for instance atom-based and structure-based) approaches is an area open for development.

ACKNOWLEDGEMENTS

The authors acknowledge Michele Ceriotti and the members of COSMO lab for useful discussion.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Bonchev, D. Chemical Graph Theory: Introduction and Fundamentals; Chemical Graph Theory Taylor & Francis, 1991.

[2] Minkin, V. I. Pure Appl. Chem. 1999, 71, 1919.

[3] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. J. Comput.-Aided Mol. Des. 2016, 30, 595.

[4] Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. J. Chem. Phys. 2018, 148, 241722.

[5] St. John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. J. Chem. Phys. 2019, 150, 234111.

[6] Weininger, D. J. Chem. Inf. Comput. Sci. 1988, 28, 31.

[7] McNaught, A. Chem. Inter. 2006, , 6.

[8] Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. Chem. Sci. 2018, 9, 513.

[9] Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. ACS Cent. Sci. 2018, 4, 268.

[10] Popova, M.; Isayev, O.; Tropsha, A. Sci. Adv. 2018, 4, eaap7885.

[11] Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. ACS Cent. Sci. 2017, 3, 1103.

[12] Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. Chem. Sci. 2018, 9, 6091.

[13] Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Chem. Mater. 2019, 31, 3564.

[14] Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Nat. Commun. 2017, 8, 15679.

[15] Calle-Vallejo, F.; Tymoczko, J.; Colic, V.; Vu, Q. H.; Pohl, M. D.; Morgenstern, K.; Loffreda, D.; Sautet, P.; Schuhmann, W.; Bandarenka, A. S. Science 2015, 350, 185.

[16] Rossi, K.; Asara, G. G.; Baletto, F. ChemPhysChem 2019, 20, 1.

[17] Rück, M.; Bandarenka, A.; Calle-Vallejo, F.; Gagliardi, A. J. Phys. Chem. Lett. 2018, 9, 4463.

[18] Melville, J. L.; Hirst, J. D. J. Chem. Inf. Model. 2007, 47, 626.

[19] Aguado-Ullate, S.; Baker, J. A.; González-González, V.; Müller, C.; Hirst, J. D.; Carbó, J. J. Catal. Sci. Technol. 2014, 4, 979.

[20] Etter, M. C.; MacDonald, J. C.; Bernstein, J. Acta Cryst. B 1990, 46, 256.

[21] Pietrucci, F.; Andreoni, W. Phys. Rev. Lett. 2011, 107, 085504.

[22] Reinhart, W. F.; Panagiotopoulos, A. Z. Soft Matter 2018, 14, 6083.

[23] Bougueroua, S.; Spezia, R.; Pezzotti, S.; Vial, S.; Quessette, F.; Barth, D.; Gaigeot, M.-P. J. Chem. Phys. 2018, 149, 184102.

[24] Cao, D.-S.; Liang, Y.-Z.; Xu, Q.-S.; Hu, Q.-N.; Zhang, L.-X.; Fu, G.-H. Chemom. Intell. Lab. Syst. 2011, 107, 106 .

[25] Glielmo, A.; Zeni, C.; De Vita, A. Phys. Rev. B 2018, 97, 184307.

[26] Willatt, M. J.; Musil, F.; Ceriotti, M. J. Chem. Phys. 2019, 150, 154110.

[27] Bartók, A. P.; Kondor, R.; Csányi, G. Phys. Rev. B 2013, 87, 184115.

[28] Behler, J.; Parrinello, M. Phys. Rev. Lett. 2007, 98, 146401.

[29] Artrith, N.; Urban, A.; Ceder, G. Phys. Rev. B 2017, 96, 014112.

[30] Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. J. Chem. Phys. 2018, 148, 241717.

[31] Glielmo, A.; Sollich, P.; De Vita, A. Phys. Rev. B 2017, 95, 214302.

[32] Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Phys. Rev. B 1983, 28, 784.

[33] Grisafi, A.; Wilkins, D. M.; Michael J Willatt, M. J.; Ceriotti, M. ArXiV 2019. ArXiv:1904.01623.

[34] Geiger, P.; Dellago, C. J. Chem. Phys. 2013, 139, 164105.

[35] Cubuk, E. D.; Schoenholz, S. S.; Rieser, J. M.; Malone, B. D.; Rottler, J.; Durian, D. J.; Kaxiras, E.; Liu, A. J. Phys. Rev. Lett. 2015, 114, 108001.

[36] De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Phys. Chem. Chem. Phys. 2016, 18, 13754.

[37] Anelli, A.; Engel, E. A.; Pickard, C. J.; Ceriotti, M. Phys. Rev. Materials 2018, 2, 103804.

[38] Wilkins, D.; Felix, M.; Ceriotti, M. Phys. Chem. Chem. Phys 2018, , 29661.

[39] Grisafi, A.; Ceriotti, M. ArXiV 2019. ArXiv:1909.04512.

[40] Bailar, J. C. J. Inorg. Nucl. Chem. 1958, 8, 165.

[41] Pinsky, M.; Avnir, D. Inorg. Chem. 1998, 37, 5575.

[42] Alvarez, S.; Alemany, P.; Casanova, D.; Cirera, J.; Llunell, M.; Avnir, D. Coord. Chem. Rev. 2005, 249, 1693 .

[43] Lechner, W.; Dellago, C. J. Chem. Phys. 2008, 129, 114707.

[44] Peters, B. J. Chem. Phys. 2009, 131, 244103.

[45] Zimmermann, N. E. R.; Vorselaars, B.; Quigley, D.; Peters, B. J. Am. Chem. Soc. 2015, 137, 13352.

[46] Lazar, E. A.; Han, J.; Srolovitz, D. J. Proc. Natl. Acad. Sci. 2015, 112, E5769.

[47] Larsen, P. M.; Schmidt, S.; Schiøtz, J. Modell. Simul. Mater. Sci. Eng. 2016, 24, 055007.

[48] Cumby, J.; Attfield, J. P. Nat. Commun. 2017, 8, 14235.

[49] Fop, S.; McCombie, K. S.; Wildman, E. J.; Skakle, J. M. S.; Mclaughlin, A. C. Chem. Commun. 2019, 55, 2127.

[50] Waroquiers, D.; Gonze, X.; Rignanese, G.-M.; Welker-Nieuwoudt, C.; Rosowski, F.; Göbel, M.; Schenk, S.; Degelmann, P.; André, R.; Glaum, R.; Hautier, G. Chem. Mater. 2017, 29, 8346.

[51] Clemens, O.; Berry, F. J.; Wright, A. J.; Knight, K. S.; Perez-Mato, J.; Igartua, J.; Slater, P. R. J. Solid State Chem. 2013, 206, 158.

[52] Balachandran, P. V.; Young, J.; Lookman, T.; Rondinelli, J. M. Nat. Commun. 2017, 8, 14282.

[53] Balachandran, P.; Benedek, N.; Rondinelli, J. Information Science for Materials Discovery and Design; Lookman, T.; Rajan, K.; Alexander, F., Eds.; Springer Verlag: Germany Springer Series in Materials Science. 2015; p. 213.

[54] Smith, J. S.; Isayev, O.; Roitberg, A. E. Chem. Sci. 2017, 8, 3192.

James Cumby received his MSc. (2010) and Ph.D. (2014) in chemistry from the University of Birmingham. Following work as a postdoctoral fellow with Prof. Paul Attfield in the Centre for Science at Extreme Conditions at the University of Edinburgh, he was appointed as lecturer in inorganic chemistry in 2019. His research focuses on combining computational and experimental techniques to develop new solid state functional materials, with a particular interest in metal oxide-fluorides materials.

Kevin Rossi received his BSc Physics from King's College London in 2014 and defended his Ph.D. Thesis in 2018 in the same university. Currently, he his a postdoctoral researcher in Prof. Michele Ceriotti Laboratory of Computational Science and Modelling (COSMO) at Ecole Polytechnique Federale de Lausanne. In his research Kevin applies multi-scale and data-driven methods to study systems and processes relevant to heterogeneous or homogeneous catalysis.