

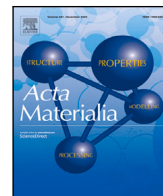


NOMATEN

This work was carried out in whole or in part within the framework of the NOMATEN Center of Excellence, supported from the European Union Horizon 2020 research and innovation programme (Grant Agreement No. 857470), the European Regional Development Fund via the Foundation for Polish Science International Research Agenda PLUS programme (Grant No. MAB PLUS/2018/8), *and the initiative of the Ministry of Science and Higher Education 'Support for the activities of Centers of Excellence established in Poland under the Horizon 2020 program' under agreement No. MEiN/2023/DIR/3795.*

This is a copy of the publication which appeared in Acta Materialia, Volume 268, 15 April 2024, 119773, published in February 2024.

DOI: 10.1016/j.actamat.2024.119773



Full length article



Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals

Dario Massa ^{a,b,*}, Efthimios Kaxiras ^c, Stefanos Papanikolaou ^a

^a NOMATEN Centre of Excellence, National Centre for Nuclear Research, Warsaw, Poland

^b IDEAS-NCBR, Warsaw, Poland

^c Department of Physics, Harvard University, Cambridge, MA, USA

ARTICLE INFO

Keywords:

Material informatics
Ab-initio
Electronic charge density
Machine learning
Density functional theory
DFT

ABSTRACT

Materials design has traditionally evolved through trial-error approaches, mainly due to the non-local relationship between microstructures and properties such as strength and toughness. We propose ‘alloy informatics’ as a machine learning based prototype predictive approach for alloys and compounds, using electron charge density profiles derived from first-principle calculations. We demonstrate this framework in the case of hydrogen interstitials in face-centred cubic crystals, showing that their differential electron charge density profiles capture crystal properties and defect-crystal interaction properties. Radial Distribution Functions (RDFs) of defect-induced differential charge density perturbations highlight the resulting screening effect, and, together with hydrogen Bader charges, strongly correlate to a large set of atomic properties of the metal species forming the bulk crystal. We observe the spontaneous emergence of classes of charge responses while coarse-graining over crystal compositions. Nudge-Elastic-Band calculations show that RDFs and charge features also connect to hydrogen migration energy barriers between interstitial sites. Unsupervised machine-learning on RDFs supports classification, unveiling compositional and configurational non-localities in the similarities of the perturbed densities. Electron charge density perturbations may be considered as bias-free descriptors for a large variety of defects.

1. Introduction

Hydrogen has, nowadays, a dual importance, both for renewable energy solutions [1,2], as well as for the fundamentals of hydrogen-embrittlement effects [3–7]. Strict criteria for hydrogen storage systems automotive applications [8] on capacities, working temperatures, reversibility of cycles, safety and price require further improvements in the deployed materials, beyond hydrides [9–11] and other solutions. For constrained optimization problems like these, the use of Machine Learning (ML) techniques appears essential for addressing the required complexity and navigating the optimal compositional space.

Indeed, in materials design and development, the role of ML [12–14] methods has been established as crucial in the context of data-driven approaches. These types of methods have evolved into a new field of computational materials design, referred to as Materials Informatics [15–19]. Existing open-access simulation and experiment datasets, have been used to develop ML models and predictions for electrocatalysts materials [20,21], carbon-capture materials [22,23], Metal Oxide Frameworks (MOFs) [24], organic materials [25,26], and many other classes, with the common aim of maximizing the screening

efficiency and materials performances. Conventionally, the foremost challenge, besides the presence of an efficient and well-tested ML algorithm, is the specification of the set of effective system descriptors.

In the context of ML approaches with atomistic input, i.e. coming from simulation data, such as density functional theory (DFT) [27–33], a large variety of descriptors have been proposed, including fixed-length feature vectors of elemental or electronic properties [34–36], as well as structural descriptors, based on rotationally and translationally invariant functions of atomic coordinates, like the Coulomb matrix [37], the atom-centred symmetry functions (ACSFs) [38], the social permutation invariant coordinates (SPRINT) [39], the smooth overlap of atomic positions (SOAP) [40] and the global minimum of root mean-square distance [41]. However, given that the number of DFT calculations is naturally limited, the challenge in selecting the appropriate set of descriptors stems from the necessity for them to effectively encompass element-specific microstructural features while also accounting for potential electron–electron correlations across various length scales.

In the search of hydrogen-based energy solutions, as well as in studies of embrittlement effects, hydrogen may be regarded as a defect

* Corresponding author at: NOMATEN Centre of Excellence, National Centre for Nuclear Research, Warsaw, Poland.

E-mail address: dario.massa@ncbj.gov.pl (D. Massa).

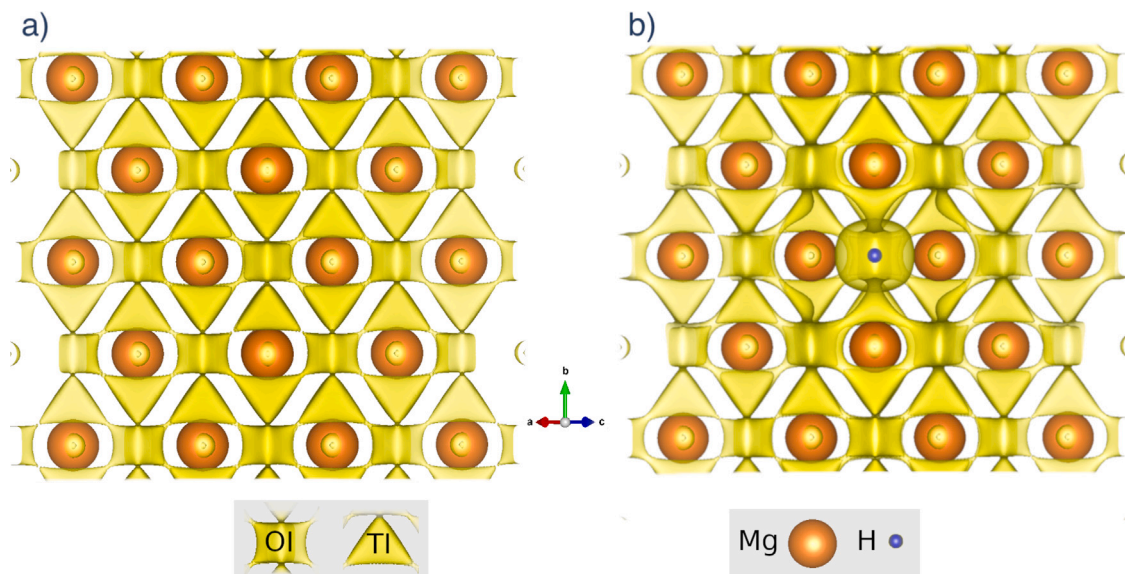


Fig. 1. Three dimensional differential electron charge accumulation isolines in a bulk Mg supercell, (a) in the absence of H and (b) in the presence of an H atom sitting at a high symmetry site. The isoline level, of $0.00183 \text{ e}/\text{bohr}^3$ in both visualizations, has been set to visibly highlight the two high-symmetry interstitial regions in which hydrogen could stably sit, namely OI and TI.

inside the host material, thus the classification of fundamental interactions among the two for constructing optimal ML descriptors represents the critical challenge. In the present work, we use the defect-induced electron charge density perturbations in order to illuminate classification features of local crystalline defects and their properties. We consider hydrogen interstitial defects in a variety of bulk face-centred cubic (FCC) metal crystals as a key demonstrative example of a high-throughput procedure [42,43], where our descriptors are constructed on the base of hydrogen differential electron charge density profiles, and of their local quantities, such as radii and extrema. In particular, in Section 3.1 we show how the building of a Radial Distribution Function (RDF) in the density fields allows for a fair comparison between the profiles, and the distinctive features of the effects of hydrogens are discussed in terms of charge density perturbations. In Section 3.3, we prove that the electron density related features show remarkable correlations with both atomic and structural features of the host bulk crystals, and that they can also be exploited to infer hydrogen mobility in the surrounding matrix. After the building of a density dataset, in Section 3.4 we perform unsupervised ML, while in Section 3.2 we investigate over the possibility of a mean-field like treatment of the density perturbations, averaging out compositional peculiarities. Basing on preliminary tests, we discuss interesting trends between the profile features and the properties of the embedding crystals, uncovering compositional relationships (among different crystals) and non-compositional ones (among different hydrogen interstitial equilibrium positions in the same crystals). The results of this study might pave the way to novel material informatics methods for prediction of defects behaviour in materials and their related effects, by interpretation of the perturbed differential charge density profiles (PChDP) in its surroundings.

2. Methods

We focus on the hydrogen-induced changes in the differential [44] electron charge density profile of an otherwise periodic crystal structure. For this aim, we develop 35 simple FCC supercells of pure metallic crystals, listed in Tab.S.(I) of the associated Supplementary Material (SM), in which a hydrogen atom is inserted. An example of such considered structures is contained in Fig. 1.

2.1. Density functional theory methods

We perform Density-Functional-Theory (DFT) simulations in the QUANTUM ESPRESSO (QE) package [45–47] of $2 \times 2 \times 2$ supercells of pure FCC metallic bulk crystals, whose list is collected in Tab.S.(I) of SM. Pseudo-Potentials for all involved atomic species are taken to be Ultra-Soft with an additional implementation of the Perdew–Burke–Ernzerhof (PBE) [48] functional.

In order to better model dispersion forces in the possible presence of weakly interacting states, the non-local exchange–correlation density functional rvv10 [49] from Vydro and Van Voorhis has been also implemented. Further, smearing [50] has been introduced to correctly deal with metals, and the calculations have been set to be spin-polarized for possible non-trivial magnetic effects. The convergence with respect to the total energy on the number of k-points, plane-waves cutoff energy and smearing spreading is checked, for each case, after a preliminary variable-cell relaxation of the pure metals, and the final equilibrium bulk structures are then obtained via a further fixed-cell relaxation with the optimal parameters. The common acceptance threshold in the variation of the total energy upon parameter change is set to 10^{-5} Ry. This process allows to ensure the correct optimization of both parameters and structures for each specific case, which might need different optimal parameters sets. The parameter sets for each atomic species are contained in Tab.S.(I) of SM. The forces and total energy convergence thresholds for ionic minimization are set to a common value of 10^{-5} a.u. and 10^{-6} a.u. respectively. The structural properties like the bulk, Young and shear moduli and the Poisson ratios of the pure optimized samples are then extracted via the THERMO_PW [51] driver of QE, and are reported in Tab.S.(II) of SM, together with the crystals lattice constants and other relevant structural information resulting from this optimization process.

In the relaxed, converged crystalline supercells, we insert one hydrogen atom, and let a relaxation minimize the forces in the system, confirming the well known two possible equilibrium interstitial positions of hydrogen in FCC structures: either as an octahedral interstitial (OI) or a tetrahedral interstitial (TI) [52].

In Fig. 1, there is a demonstration of differential electron charge density isolines in the absence and presence of H in a Mg supercell. The accumulation isoline level has been set to the same one in both

Table 1

Density-related properties promoted as describing features of hydrogen interaction in different materials. (M) stands for Metal and (H) stands for Hydrogen.

Feature	Description
Ch(M) ex.(e ⁻)	exchanged charge from the M atom
Ch(H) ex.(e ⁻)	exchanged charge from the H atom
R(M)(Å)	Bader radius of the M charge
R(H)(Å)	Bader radius of the H charge
V(M)(Å ³)	Bader volume of the M charge
V(H)(Å ³)	Bader volume of the H charge
RDF peak width	half-height width of the RDF peak
RDF MAX	maximum of the RDF peak

cases, and in such a way to visibly highlight the two high-symmetry states interstitial regions in which hydrogen could stably sit.

We extract the PChDP around hydrogen atoms, by subtracting the differential density of pure bulk structures from the ones of defected samples, both for OI and TI cases. At the same time, for all samples and configurations, the electron (pseudo-)charge densities and the all-electron ones are also stored and used in Bader charge analysis [53]; the latter provides useful information regarding Bader charge exchanges, volumes and radii [54]. In Tab.S.(III) and Tab.S.(IV) of SM, these quantities are documented for each studied crystal. In particular, for building a dataset with atomic, structural and density properties, we consider a set of elemental properties from the Mendeleev package, a Python resource [55], and we maintain Bader Analysis [53] quantities related to the hydrogen atom and to one of its first nearest neighbour. Technically, in an FCC crystal in which hydrogen lives as OI, there are six nearest neighbours, but the extracted Bader information is nearly identical for symmetry. For simplicity and consistency, once the choice of which metal atom should be selected for its Bader properties is done in the OI case, the same metal atom is chosen for the TI case as well.

In order to compute the hydrogen binding energies (BEs) in different crystals, we consider the following formula

$$E_{BE} = E_{MH} - E_M - \frac{E_{H2}}{2} \quad (1)$$

where E_{MH} is the total energy of the bulk supercell with the hydrogen interstitial atom, E_M the total energy of the supercell without hydrogen, and E_{H2} the energy resulting from a fixed-cell relaxation of an hydrogen molecule in a sufficiently large cubic simulation box ($a = 24\text{Å}$). The BE is computed with the same QE parameters used for the specific metal bulk system each time in consideration, for computational consistency.

2.2. Isolines and materials' comparisons

In order to access to a more complete description of the effects of a systematic change in composition of the embedding FCC matrices, and the dependence on the different hydrogen interstitial positions, we investigate the contours of the PChDP around hydrogen. All the density grids are brought to an identical resolution by using mp-pyRho[56]. Then, identifying a unique radius R in the unit cell to which each value of the density profile can be associated, a Radial Distribution Function (RDF) is calculated [57], representing the basis of our comparison between different profiles. The task is parallelized using the Joblib Python package. Finally, shifting the obtained signals by the hydrogen position in the grid is a key step for efficiently comparing the profiles. In Table 1 are reported the relevant density-related properties which have been further considered.

2.3. Validation strategies and procedures

2.3.1. Diffusive properties

The key aspects related to hydrogen for industrial applications, especially for renewable energy such as hydrogen storage or catalysis of energy reactions, ultimately rely on its diffusive characteristics

in different materials [58–61]. In order to identify validation routes based on the natural relationship between the extension of a charge profile and its mobility in the surrounding environment, we utilize DFT Nudge-Elastic-Band (NEB) calculations to quantify energy barriers for hydrogen migration along highly-symmetric sites in bulk metals. In particular, $2 \times 2 \times 2$ supercells with one hydrogen atom have been considered as starting and ending images, and the total number of points to discretize the path has been set to 14, with quasi-Newton Broyden optimization scheme and optimization step length of 1.0 a.u.. The threshold controlling the error (the norm of the force orthogonal to the path) in the climbing-image simulation is set to 0.05 eV/Å.

2.3.2. Size effects in charge density calculations

The charge density perturbations due to an interstitial hydrogen defect shall display size effects, that may gain crucial importance for the possibility of undesired artifacts from spurious interactions in small supercells. As sanity checks, in Fig.S.(1) and Tab.S.(V) of the SM we report the comparison of a three dimensional perturbation induced by an hydrogen interstitial defect for a $2 \times 2 \times 2$ and $3 \times 3 \times 3$ Mo supercell, proving that in typical metals, and thus, for the aims of this study, the expected size effects are negligible.

2.3.3. Effects of different crystalline symmetries in charge density considerations

For the applicability of the proposed method, we mainly consider the important example of FCC pure bulk crystals, materials that are predominantly metals. Clearly, not all the considered elements form FCC crystals at Standard Temperature and Pressure (STP) conditions, but we opted for a single-phase study for the sake of a simpler comparison among the profiles, avoiding possible differences arising from different crystalline backgrounds and interstitial positions which could complicate the origin of charge density fluctuations. However, the methods discussed in this manuscript are general and shall become applicable to other crystalline structures. For this purpose, the SM contain examples of hydrogen-induced PChDP for non-metallic HCP supercells, including C, Si and Hg, and we promote specific relevant considerations and comparisons with the results presented in the main text.

2.4. Machine learning

2.4.1. Principal Component Analysis (PCA)

In the present work, we follow the Principal Component Analysis [62,63] method, probably the most popular choice for developing unsupervised learning, allowing the simultaneous preservation of as much information (variability) as possible. Let us consider a dataset represented by a $n \times p$ matrix, \mathbf{D} , where n is the number of samples and p is the number of features in the dataset. If we define a linear combination of features as

$$\sum_{j=1}^p c_j \mathbf{d}_j = \mathbf{Dc} \quad (2)$$

where \mathbf{c} is a p -dimensional vector of constants, the aim is to search the combination such that its variance is maximized, which means maximizing the quadratic form

$$\text{var}(\mathbf{Dc}) = \mathbf{c}^T \mathbf{S} \mathbf{c} \quad (3)$$

where \mathbf{S} is the $p \times p$ covariance matrix. Requiring with a Lagrange multiplier the vectors of constants to have unit norm, the actual problem to solve becomes the following

$$\max[\mathbf{c}^T \mathbf{S} \mathbf{c} - \lambda(\mathbf{c}^T \mathbf{c} - 1)] \quad (4)$$

and the consequent differentiation with respect to the vector of constants leads to an eigenvalue problem which is at the base of PCA

$$\mathbf{S} \mathbf{c} = \lambda \mathbf{c} \quad (5)$$

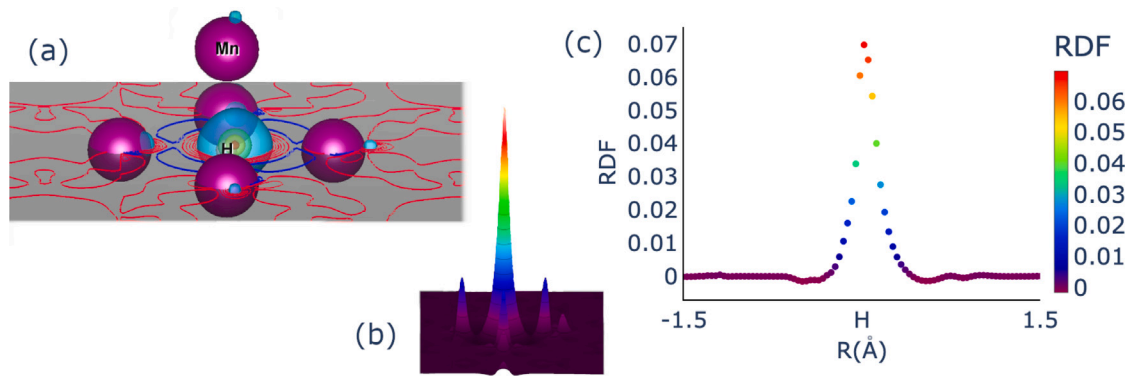


Fig. 2. (a) Three dimensional visualization of the PChDP of Mn (OI case), resulting from the subtraction of the differential profiles of the pure crystals from the defected ones; multiple 3d-isolines are highlighted, starting from the highly positive (accumulation) in red (0.07 e/bohr³), to positive in yellow and green (0.05 and 0.03 e/bohr³ respectively), to faintly positive, in blue (0.009 e/bohr³). On the highlighted grey sectional plane, some accumulation (red) and depletion (blue) isolines are represented, and (b) displays the corresponding bird-eye visualization of the perturbation, with a colour gradient for the charge value mapping the isolines values displayed in the 3D visualization panel, ranging from -0.0043 e/bohr³ to 0.0759 e/bohr³, and with out-of-plane extension basing on the intensity. (c) The RDF profile obtained from the PChDP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Therefore, λ are the eigenvalues of the covariance matrix \mathbf{S} corresponding to the unit norm eigenvectors \mathbf{c} . Being \mathbf{S} a squared real symmetric matrix, it has exactly p eigenvalues, corresponding to eigenvectors \mathbf{a}_k that are the solution of the problem of finding p new linear uncorrelated combinations (Principal Components, PCs) $d_{\mathbf{a}_k} = \sum_{j=1}^p a_{jk} d_j$ that maximize the variance [64]. In a geometric interpretation of PCA, we can interpret the eigenvalues of the covariance matrix as associated to the magnitude of the variance captured by the corresponding PC (axis) in the multivariate cloud space, and the eigenvectors as the orientation of the PC (axis). The so-called loadings, eigenvectors scaled up by variances, are the carriers of the combined information on the captured variance and direction of the PCs, and, representing the covariances between the latter and the original dataset variables, tell us how much of the variation in a variable is explained by the component.

2.4.2. K-Means clustering

Unsupervised learning can be accomplished by combining PCA and K-Means clustering. The K-Means algorithm was proposed by Lloyd in 1982 [65] to solve the problem of finding the minimum total squared distance of each point $x \in X \subset \mathbb{R}^d$ from its closest centre $m \in M$ of cluster $c \in C$.

$$\phi_X = \sum_{x \in X} \min_{m \in M} \|x - m\|^2 \quad (6)$$

A survey from ten years ago [66] reports it already was, and probably still is, the most used clustering algorithm. The problem is solved in the following way:

- Define arbitrary (typically random) centres $m_1^{(t=0)} \dots m_k^{(t=0)}$;
- Define a cluster c_i as the set of points closer to m_i than m_j , for each $j \neq i$;
- Define the new k-centres (k-means) as $m_i^{(t+1)} = \frac{1}{|c_i|} \sum_{x \in c_i} x$
- Repeat the previous two steps until the set of means is converged

Even though the target total squared distance is a monotonic decreasing function of the number of clusters and, since there are k^n possible clusters, the algorithm is proved to converge, bad clustering can be obtained without a proper initialization. For this reason, Arthur and Vassilvitski introduced a seeding algorithm [67], K-Means++, able to outperform the results of usual K-Means just by weighting the selected data-points according to their squared distance from the already randomly selected centre.

A part from a proper initialization, whenever no ground truth is available about the number of cluster in which the data is expected to classify in, there is the need of finding the optimal number of clusters. A variety of methods is available in order to perform the cluster number

analysis, and we here focus on the Elbow analysis [68], based on the sum of squared distances [69] in Eq. (6): when its decrease with the number of clusters starts to be negligible, the optimal number is found, above which overfitting occurs.

3. Results

3.1. The RDFs

In order to better explain the building and meaning of the OI and TI RDF profiles, let us focus on one specific example, the case of an hydrogen atom as an octahedral interstitial defect in a Mn supercell, shown in Fig. 2.

The PChDP isolines can be visualized in the three dimensional structure of the supercell, as in panel (a): the PChDP has its positive maximum (accumulation) close to the hydrogen atom, and fades out to a zero level oscillation with increasing distance from it, as the progressively weaker isolines report. In panel (b) we remove one dimension from the visualization, in order to appreciate the perturbation on a plane that crosses the positive maximum. The colour gradient for the charge value is mapping the isolines values displayed in the 3D visualization panel, but is also strictly related to the RDF peak height of panel (c): it leads to the removal of the orienteering dependence and is built as explained in Section 2. If one now repeats the extraction for both the OI and TI cases of different samples, as described in Section 2.2, obtains the result shown in Fig.S.(7) of the SM. For completeness, some relevant quantities from the plots are reported in Table(III) and Table(IV) of SM, which include peaks heights and widths.

It is important to notice how even at this level, without the support of any ML model, one is able to make some compositional-based comparison between the investigated cases, by looking at the peaks properties. An intuitive but strong relation providing a first insight on the presented curves could be found in the charge densities themselves. If one considers the obtained maxima values for the RDF profiles and the properties of the charge density profiles coming from Bader charge analysis, as in Fig. 3, good linear relations are found, especially with the hydrogen Bader charges. The finding suggests the tendency of having higher RDF peaks for larger accumulated charges on hydrogen, which stands as reasonable thinking about the RDF perturbation as a representative of the charge density response (screening) to the added impurity, and given that the RDF profiles themselves are built on the hydrogen PChDPs. However, a clearer picture requires the exploration of correlations with the entire spectrum of atomic-structural and density properties, as will be proposed in Section 3.3. Here we would like to stress that the RDF itself is providing an isotropic measure

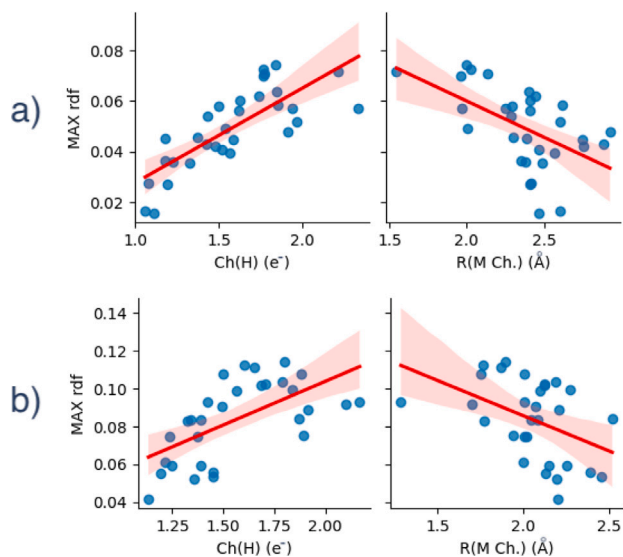


Fig. 3. Linear fits and 95% confidence regression intervals on the relation among RDF peaks properties (MAX rdf) and the charge density ones of the hydrogen (H) and metal atoms (M), both for OI (a) and TI (b) configurations.

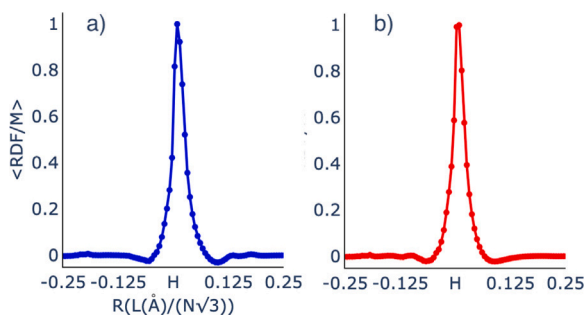


Fig. 4. RDFs averaged over all the available configurations (Eq. (7)), for both OI (a) and TI (b) states. The axes share the same labels.

of the PChDP in the crystal due to the insertion of the H atom, and its comparison for different metallic crystals can clearly provide a measure of similarity of the charge response. For this reason, it can be considered as the descriptor of the perturbed profiles, being a unique signal and fingerprint of the defect-crystal interaction.

In the next paragraph, we question the possibility of a unique average descriptor, on the basis of which a first classification of the system-specific defect-induced perturbations can be demonstrated.

3.2. Composition specific fluctuations from an average oscillation

It is of interest the possibility of considering a H-universal RDF curve for our problem, capable of grasping the average behaviour of all the system specific charge density responses that we have observed, assuming crystal composition as a degree of freedom. In order to do so, one could consider the average over N different bulk crystals

$$\left\langle \frac{RDF_i}{M_i} \right\rangle_{i=1..N}, \quad M_i = \max(RDF_i) \quad (7)$$

which is proposed in Fig. 4. The resulting averaged curves of OI and TI cases are very similar, a part from an inverted symmetry noticeable in the short-range rippling of the profiles.

In order to be able now to quantify the capabilities of a coarse-grained perturbation of grasping the peculiarities of the system specific ones, let us consider the deviations from the average

$$\frac{RDF_i}{M_i} - \left\langle \frac{RDF_i}{M_i} \right\rangle_{i=1..N}, \quad M_i = \max(RDF_i) \quad (8)$$

which are reported in Fig. 5, both for OI and TI states.

Interestingly, again without any ML support, the deviations can be easily grouped in three behavioural classes basing on the minima and maxima positions, which also show a conservation in their overall characteristics going from the OI to TI state, a part, as for the averaged RDFs, from an inverted symmetry along the axis of distances from their centre (hydrogen). As it can be noticed from the periodic table plots, the majority of the group members are conserved going from the OI to TI state, except for few changes involving Na, Mo, Rh, In, Sn, Os and Ir.

Going from Group 1 to Group 2, which together include almost all the non-noble metals, the local deviations from the average charge perturbation gain two global minima which are almost symmetric with respect to their centre: thinking in terms of the original characteristics of the signals appearing in the operation of Eq. (7), this should imply a change in the width of the crystal specific one, with respect to the averaged perturbation. Interestingly, we notice how noble-like metals, namely Au, Pd, Pt, Rh and Ir, are classified inside Group 3 (OI), and are characterized by a considerable increase in the number of maxima and minima of the presented deviations. This special class of metals is generally known for their stability, resistance to corrosion and oxidation in moist air, as well as being effective catalysts for the hydrogen evolution reaction (HER) [70]. Other transition metals from Group 2, such as Ti, V and Cr, when combined to form an high-entropy alloy (HEA), have been proved to absorb hydrogen at room temperature and with moderate equilibrium pressures for the dihydride formation [71].

Even though verifying the actual catalytic activity or hydrogen storage performances of the proposed crystals, or their combinations, in the presence of hydrogen molecules is beyond the scope of the present work, we believe these observations inspire ourselves, and possibly the readers, for further work in this direction. It is worth to point out that the proposed element-specific deviations are a sophisticated form of characterization of the metal-hydrogen interaction, inheriting the richness of the presented descriptors, and delivering a bias-free classification of the local peculiarities of the charge density responses due to characteristic atomic (therefore, electronic) environments. One may argue the inter-classes variations in the number of stationary points to be related to a change in the energy of the electrons providing the screening of the impurity: a longer rippling profile shall be associated to lower energies, and indeed, as we will show in the next sections, the accumulated charge on H is significantly minimized in the noble metals.

In the next paragraph, the capabilities of our descriptors to correlate with the physical properties of the systems are quantitatively investigated, and will shed light over the intertwined relations driving the presented classification. In Section 3.4, an alternative and more traditional classification method will be presented basing on a set of density-related features, investigating their actual influence on the samples distribution.

3.3. The density dataset: correlations

In order to prove the key role of the PChDP as effective descriptors, a dataset has been built on the base of the density information, from the Bader charge and RDF analysis, as well as on the structural and atomic one, collected for each metallic crystal here considered. The Pearson correlation matrix for the dataset can guide our attention on the most relevant features, and, given that the relationships between atomic and structural properties are well known, in Fig. 6 we focus on some specific sub-blocks.

Let us focus first on the density related features. Some expected behaviour is obtained: (i) M bader charge and H bader charge exchanges show a remarkable positive correlation (0.7 in scale), implying the exchange from M to H (Tab.S.(III,IV) in S.M.) (ii) the larger is the number of valence electrons of the metals (M), the smaller the charge

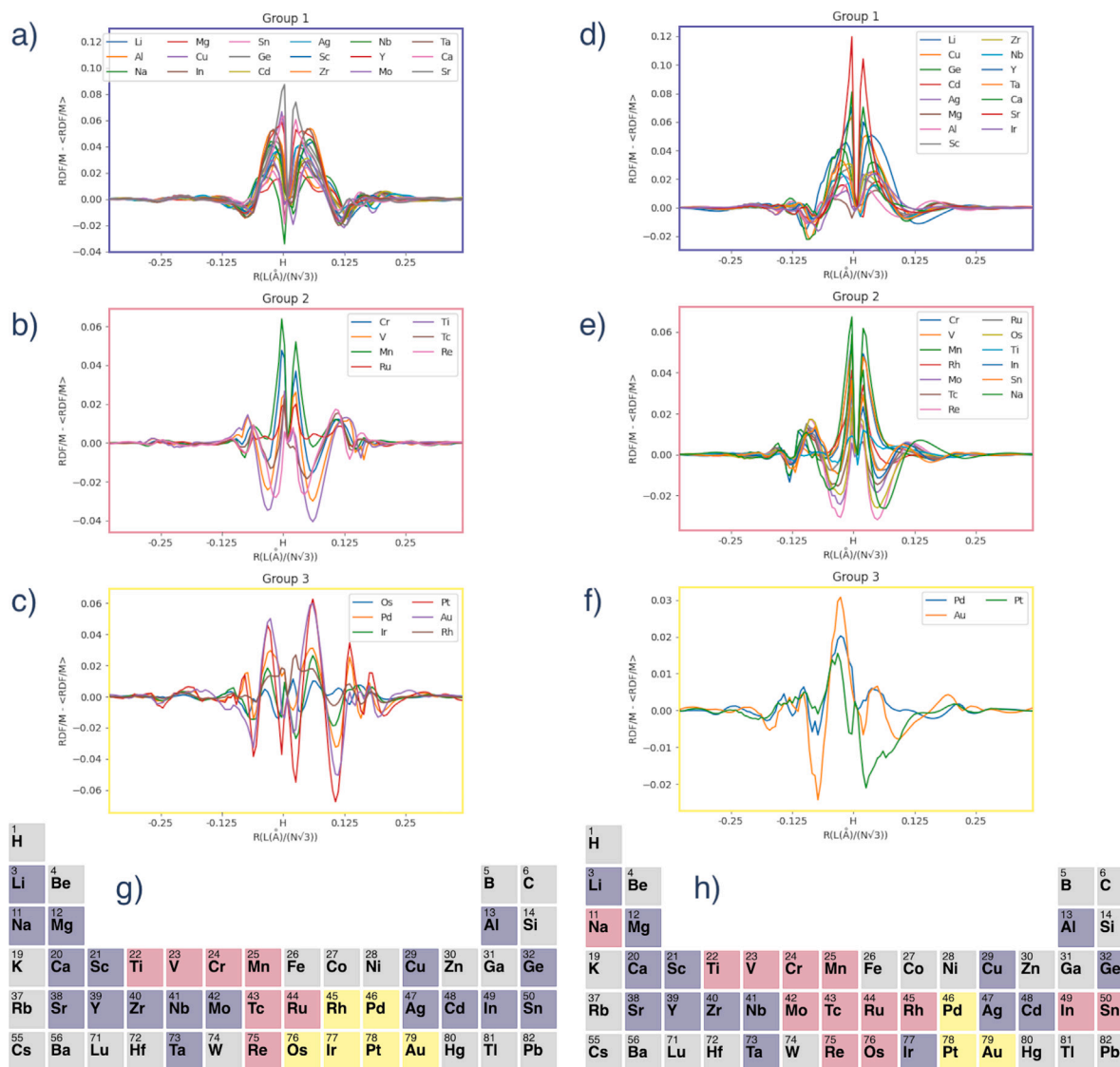


Fig. 5. Crystal-specific oscillations from the rescaled and averaged RDF (Eq. (8)), both for OI (a: Group1, b: Group2, c: Group3) and TI (d: Group1, e: Group2, f: Group3) states. The three behavioural groups are highlighted by a coloured plot frame, which helps highlighting them in the periodic tables below (g: OI, h: TI). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

it exchanges with (to) hydrogen (H); (iii) the two definitions of bader charge extension in terms of radius and volume are largely correlated; (iv) low correlation is present between valence and radius of the M charges, due the expected correlation of the latter with the definition of the atomic radius [72], and therefore the periodic behaviour along the periodic table. Moreover, the maximum value and the width of the computed RDF peaks in the grid of the hydrogen (H) PChDP is strongly anti-correlated with the Bader charge extension of the metal (M) both in terms of radius [73] (-0.6 and -0.5 respectively in scale) and volume (-0.5 and -0.7 respectively in scale), and hydrogen charges properties are affecting the RDF peaks in an overall similar way. From the similar behaviour of M and H charge extensions, the similar influence of the latter on the RDF peaks properties can now be understood.

Let us now consider the correlations between the density related features and the atomic features of the involved metals. Interestingly, the RDF peaks widths show strong anti-correlations with the metal atomic radius, atomic volume and any other definition of bonding radius, and the reason lies in the close relationship that the RDF peaks have with the metal and hydrogen charge features: the bader charge extensions of the metals are fully correlated with the definitions of

atomic volume and radius, and very tight relationships also hold with hydrogen charge extensions.

Coming back to Fig. 5, the reader can now understand the sensitivity of the descriptors properties to changes in the host crystalline matrix composition, which is at the base of the previously emerged classification: for example, discrepancies in widths between crystal-specific and a-specific perturbations (averaged) drive the oscillatory behaviour of their difference, and give meaning both to an average screening, as the result of average atomic properties, as well to interstitial hydrogen as a probe of the composition-specific properties, through the induced electronic perturbation. It is also interesting to notice how non-negligible correlations are also found with larger-scale properties like melting points, heat of formation, boiling point.

Concerning the correlations between the density related features and the structural ones [74], again the PChDP information grasps very efficiently the properties of the host crystals: the Poisson ratio and shear modulus are both encoded in the assumed hydrogen bader charge with a strong anti-correlation (-0.6 in scale), while the RDF peaks properties are closely related to the lattice constants. Visualizations of the discussed intertwined relations of atomic and structural properties with the density descriptors are shown in the example of Fig. 7.

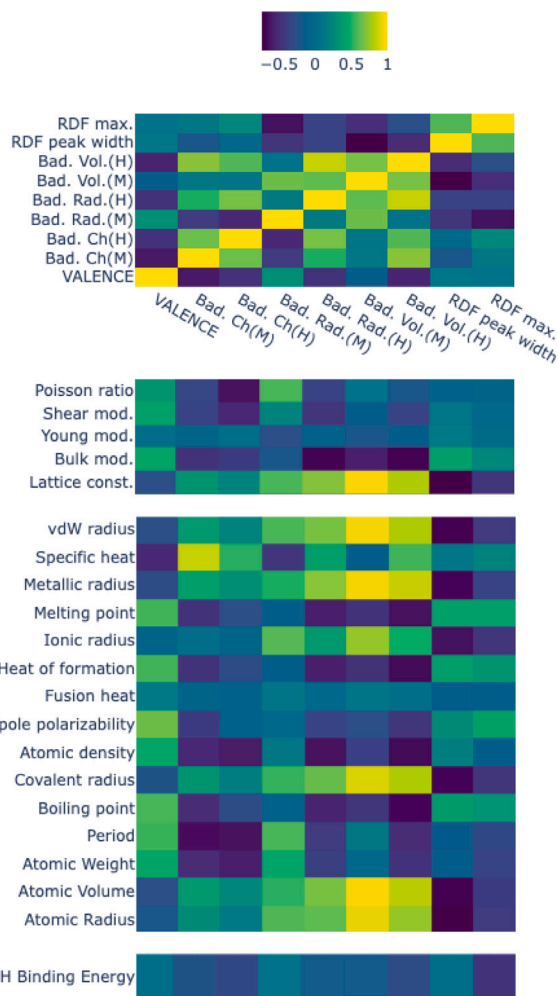


Fig. 6. Correlation matrix between the density-related features (x-axis) with themselves (higher y-block), with the structural (second y-block) and atomic (third y-block) features, and with the computed hydrogen binding energy (lower y-block).

The remarkable correlations that Fig. 6 highlights among the hydrogen BEs and charge related properties, like the RDFs maxima or the hydrogen Bader charges, have led the focus over this quantity and its variations among the different crystals, and corroborates the thesis of effectiveness of the proposed electron density fields as descriptors of the system interactions.

Fig. 8 reports the computed BE values on periodic tables for both OI and TI states. Overall, a part from few outliers, the binding tends to gradually increase along the periods, with negative values among the alkali, alkaline earth and early transition-metals, where it crosses the zero level to become positive among the late transition-metals, increasing up to the crystallogens. It is to notice that the possibility of obtaining composition maps that could drive the design of systems basing on key properties like the hydrogen binding energy is highly desirable in the context of materials for hydrogen storage: the U.S. D.O.E. has fixed indeed a specific range of interest for this value [75] in hydrogen applications. However, a reliable descriptor is needed to capture the composition induced variations, basing on its underlying power to encode local environments and interactions. In Fig.S.(8) of SM, we propose a visualization of the relation between the hydrogen Bader charges and its binding energy, which is found to be almost linear.

As a last point of this discussion, we want to prove that charge density can help in directly accessing also to the diffusion properties

Table 2

Results from the Nudge-Elastic-Band analysis of the energy barrier for the hydrogen migration along a path in different crystals. For the OI-TI path, two values are present because the barrier evaluation is different depending on whether is computed relatively to the OI (first value) or TI (second value) extremes of the path.

	OI-OI E(eV)	OI-TI E(eV)
Al	0.369	0.085 , 0.122
Mn	1.241	0.701 , 0.291

of the hydrogen atom in the different crystals, by looking for a relationship between its migration barriers and charge spatial extensions. In order to prove it, following the procedure explained in Section 2.3.1, Nudge-Elastic-Band calculations have been applied for two cases: Al (high volume of hydrogen Bader charge), Mn (low volume of hydrogen Bader charge).

The results in Fig. 9 and Table 2 show a much smaller energy barrier for the high charge volume case (in Al) with respect to the low charge volume one (in Mn). The full migration energy barrier profiles along the proposed paths also show a similar behaviour in the transition states: the favourite path is OI-TI, as also highlighted in other works for the case of Al [76]. The results also lead to an interesting comparison with a specific result in the literature, which, to our opinion, makes both findings more complete: L. Messina et al. [77] have taken into account a complementary approach for the investigation of the diffusion of impurities in bulk FCC iron; instead of systematically changing the species forming the pure FCC bulk while keeping the same impurity, as done in the present work, they have been keeping the same crystal, systematically varying the interstitial impurity. Even if hydrogen was not involved, among their findings is a counter-intuitive but largely justified correlation between the migration barriers of a solute in iron and the solute compressibility: the larger solutes diffuse faster because of lower migration barriers (a part from P and Si, in their study); moreover, the migration barriers (diffusion coefficients) are displaying a parabolic trend for 4d and 5d elements and an M-shaped (W-shaped) trend for the 3d ones.

In the present problem of systematically changing the hosting matrix around interstitial hydrogen, the extracted density information (i.e. its Bader charge volume) is closely related to the mobility of the interstitial in the system and to the compressibility of the latter, in its pure conditions: the larger volumes (smaller transition barriers) are found for larger compressibilities of the pure hosting crystals, in systems where the number of valence electrons of the metals tend to be smaller.

After having assessed the strong descriptive power of the electron density features in grasping a wide spectrum of properties of the systems of interest, in the next section we restrict to the former features and apply unsupervised ML techniques to test their ability to uncover hidden patterns and behavioural classes.

3.4. Unsupervised Machine Learning on a pure density dataset

Let us consider the application of unsupervised ML techniques to the previously presented dataset, in order to possibly reveal hidden patterns in the data that can lead to deeper understanding and characterization. At this stage, we report the result of the analysis on the pure density features dataset. As explained in Section 2.4.1, PCA is a powerful procedure to reduce the dimensionality of the dataset down to its most relevant components. The histogram in Fig.S.(2) of SM provides the explained variance ratios by each of the resulting principal components: the first two, are able to account for around 70% of the total variance, but, being the result of a transformation of the original dataset components, the individual role of the original features is not clear at this stage.

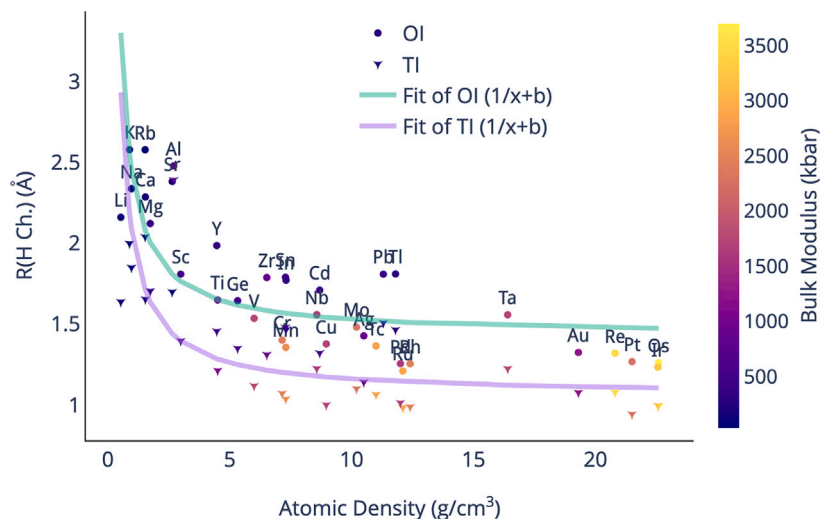


Fig. 7. Plots of specific relationships involving atomic (atomic density), structural (bulk modulus) and electron density (H Bader radius) features for both OI and TI samples, with the superimposed visualization of fitted trend-lines.

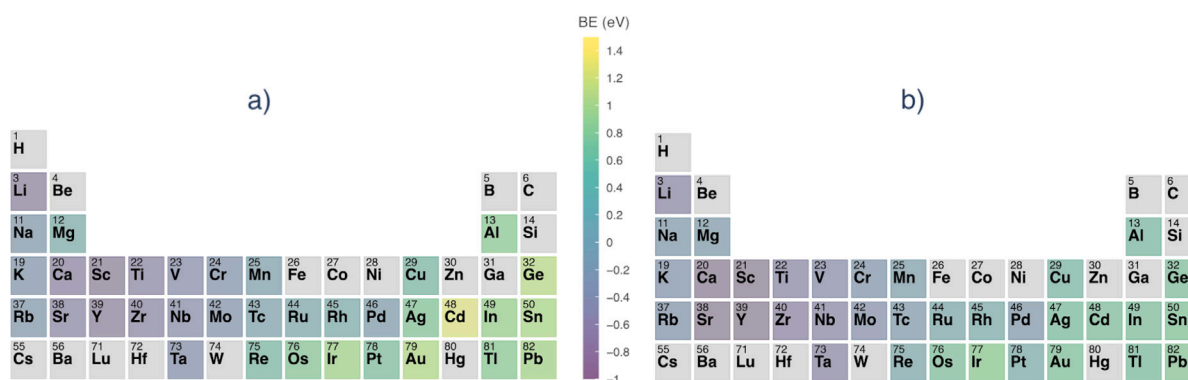


Fig. 8. Periodic table plots of the hydrogen binding energy (BE) in the OI (a) and TI (b) states of the considered FCC crystals. The colorscales are unified to the one of the single colorbar displayed. The bulk crystals that have not been considered in this study have been either removed (lanthanoids, actinoids, groups 15–18, period 7) or left in grey. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

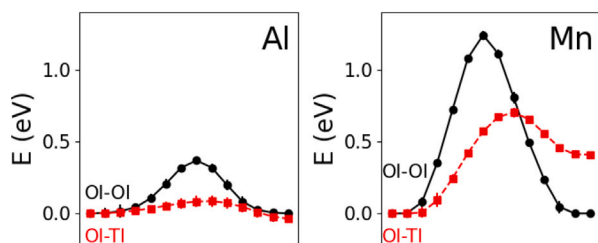


Fig. 9. Hydrogen migration energy barriers in Al (more extended Bader charge volumes) and in Mn (less extended). The barriers are the result of climbing-image NEB calculations, as explained in Section 2.3.1.

In order to try to understand whether an overall classification could emerge in the form of behavioural classes among the investigated samples, clustering methods can be applied. Since we know that the samples are characterized by two different hydrogen interstitial positions in the bulk host crystals, namely OI and TI, the most naive way of tentative clustering could be to highlight the two in the reduced PCA space and evaluate their separation, as shown in Fig. 10. As it can be noticed, the two clusters have a similar distribution in the first two principal components space, but slightly separated. The reason for the shifting lies in the electron density information, the only one differentiating the OI and TI samples at this stage. However, resorting

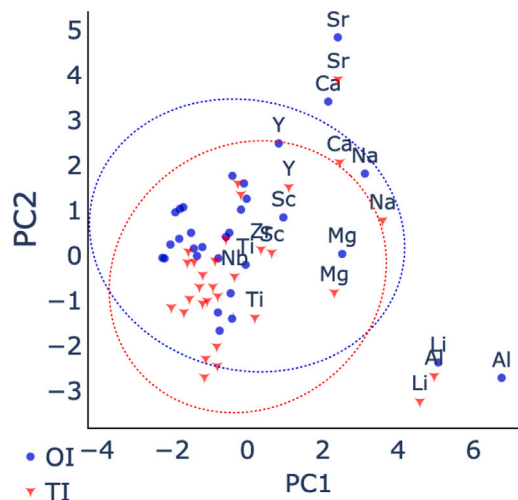


Fig. 10. Visualization of OI and TI samples in the first two-PCA components space, together with the corresponding 95% confidence ellipses. Labelling has been limited to samples out of the crowded region, in order to ease the reading of the plot.

to separation in intuitive classes does not necessarily help with the emergence of meaningful local clusters in our dataset, or with the

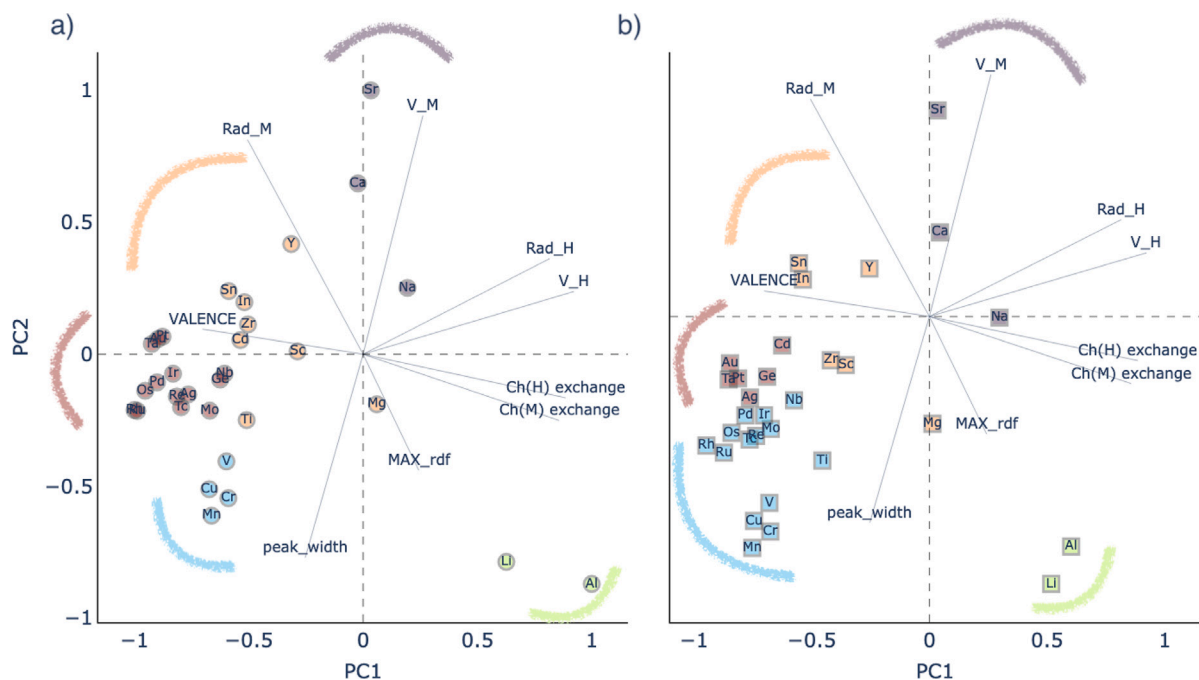


Fig. 11. The biplot for the first and second principal components, both in OI (a) and TI (b) configurations. The plot is a representation of the loadings on the two-components subspace, allowing to recognize the influence of each feature on the components, and thus their driving on the clusters distribution. In order to allow for a comprehensive visualization of the biplot, the PCs are rescaled to the $[-1, 1]$ interval. The colours for the scattered dots come from the K-Means clustering labels for the optimal number of clusters, and coloured arcs highlight the plane sectors where the corresponding clusters are found. The loadings values are presented in the Supplementary Material in Tab.S.(VI). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

understanding of the role of each specific electron density feature in driving the samples distributions.

In absence of prior knowledge on the number of classes that our dataset could distribute among, the K-Means method, introduced in Section 2.4.2, is among the ones that can be adopted to perform cluster analysis on the reduced space of the principal components. A detailed evaluation of the optimal number of clusters should be performed in the absence of ground truth, and, as shown in Fig.S.(3) of SM, the Elbow method indicates $k = 5$ as the number of clusters to optimally group our data.

A quantification of the influence of each feature is necessary and can be accessed through the so called *biplots*, in which the loadings are displayed in the reduced spaces of the principal components. As discussed in Section 2.4.1, the loadings represent the covariances between the PCs and the original variables and therefore are informative on the variables variation explained by the components. In the biplots of Fig. 11, the clusters have been highlighted following the optimal K-Means clustering previously analysed, in order to be able to associate the influence of the original variables also on the clusters distribution, besides the overall one. As it can be noticed, out of the optimal clusters, some are noticeably smaller: the one of Al and Li, i.e. maximizing the charge exchange from M (in terms of a loss) to H (in terms of a gain), or the one of Sr, Ca and Na, maximizing the measures of the metal charge extensions, minimizing at the same time the RDF peak properties. Noticeably, transition metals form three close clusters, all in favour of minimizing the amount of charge exchanged to hydrogen, and so its charge extensions. As seen in the previous subsection, charge extensions are related to migration energy barriers between high-symmetry interstitial states, and this analysis represents an example of direct access to configurational and compositional subspaces exploration for tuning the defect influence on the electron density local landscape, which, as shown, grasps a large variety of physical properties of the host matrix, and of the defect-host interaction, as we believe a general effective descriptor should do.

4. Final remarks

Through a detailed correlation analysis involving the properties of the ab-initio defect-induced charge density perturbations, we prove the ability of the latter to effectively grasp atomic and structural properties of their host metal systems. The expected strong correlations between the induced charge perturbation profiles and the underlying captured atomic charge properties, suggests the role of the latter in the long-bridging abilities of the former with a wide spectrum of crystal properties. The power of the proposed approach lies in its simplicity and generality: any defecting or functionalization of a system introduces a perturbation in the charge density profile, which, embodying the guest-host interaction, naturally inherits features of both on a pure density basis. Further reasoning comes from the brief analysis of the correlation between the hydrogen charge extensions and its transition energy barriers, for candidate systems in which the former quantity presented large differences: where hydrogen displays larger charge extensions, there are found smaller energy barriers for transitioning along different paths. This interesting result shall be further investigated, but clearly defines a strategy of predicting defects mobility just on a density-based argument. The power of the density related features in driving the distribution of the samples in a classification task has been addressed, in turn highlighting the importance of the density-aided classification task itself in studies of defects. The resulting maps, containing behavioural classes spanning different compositions and configurations, can effectively drive material searches basing on maximization/minimization of density properties, knowing their quantified strong influence on the “phenotypical” atomic/structural properties, being the density the well known associated “genotype”, in this biological analogue. A detailed analysis of crystal-specific charge density response variations from a compositionally coarse-grained one has also been performed, leading to the spontaneous emergence of behavioural groups showing the same oscillatory peculiarities. Further investigation is needed in order to fully correlate the gradual changes in the charge responses of the different crystals with complex phenomena like catalysis or storage performances. However, the power of the

approach and of the new proposed density-based descriptors gives hope and inspiration for deeper and larger scale investigations along these important directions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting the findings of this study are available upon reasonable request. Please contact the corresponding author for access to the data.

Acknowledgements

We acknowledge support from the European Union Horizon 2020 research and innovation program under grant agreement no. 857470 and from the European Regional Development Fund via the Foundation for Polish Science International Research Agenda PLUS program grant No. MAB PLUS/2018/8. We also acknowledge fruitful conversations with Prof. Ju Li, as well as the MISTI grant, funded by Lockheed-Martin, bridging the contact.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.actamat.2024.119773>.

References

- J.O. Abe, A.P.I. Popoola, E. Ajenifuja, O.M. Popoola, Hydrogen energy, economy and storage: Review and recommendation, *Int. J. Hydrogen Energy* 44 (29) (2019) 15072–15086.
- R. Moradi, K.M. Groth, Hydrogen storage and delivery: Review of the state of the art technologies and risk and reliability analysis, *Int. J. Hydrogen Energy* 44 (23) (2019) 12254–12269.
- X. Li, X. Ma, J. Zhang, E. Akiyama, Y. Wang, X. Song, Review of hydrogen embrittlement in metals: hydrogen diffusion, hydrogen characterization, hydrogen embrittlement mechanism and prevention, *Acta Metall. Sin. (English Letters)* 33 (2020) 759–773.
- P. Gong, A. Turk, J. Nutter, F. Yu, B. Wynne, P. Rivera-Diaz-del Castillo, W.M. Rainforth, Hydrogen embrittlement mechanisms in advanced high strength steel, *Acta Mater.* 223 (2022) 117488.
- G. Lu, E. Kaxiras, Hydrogen embrittlement of aluminum: the crucial role of vacancies, *Phys. Rev. Lett.* 94 (15) (2005) 155501.
- M.B. Djukic, G.M. Bakic, V. Sijacki Zeravic, A. Sedmak, B. Rajicic, The synergistic action and interplay of hydrogen embrittlement mechanisms in steels and iron: Localized plasticity and decohesion, *Eng. Fract. Mech.* 216 (2019) 106528.
- V.A. Polyanskiy, A.K. Belyaev, *Advances in Hydrogen Embrittlement Study*, Springer, 2021.
- US DOE, Target explanation document: onboard hydrogen storage for light-duty fuel cell vehicles, https://www.energy.gov/sites/prod/files/2017/05/f34/fcto_targets_onboard_hydro_storage_explanation.pdf.
- Jose Bellosta von Colbe, Jose-Ramón Ares, Jussara Barale, Marcello Baricco, Craig Buckley, Giovanni Capurso, Noris Gallandat, David M. Grant, Matylda N. Guzik, Isaac Jacob, Emil H. Jensen, Torben Jensen, Julian Jepsen, Thomas Klassen, Mykhaylo V. Lototsky, Kandavel Manickam, Amelia Montone, Julian Puszkil, Sabrina Sartori, Drew A. Sheppard, Alastair Stuart, Gavin Walker, Colin J. Webb, Heena Yang, Volodymyr Yartys, Andreas Züttel, Martin Dornheim, Application of hydrides in hydrogen storage and compression: Achievements, outlook and perspectives, *Int. J. Hydrogen Energy* 44 (15) (2019) 7780–7808.
- Shin-ichi Orimo, Yuko Nakamori, Jennifer R. Eliseo, Andreas Züttel, Craig M. Jensen, Complex hydrides for hydrogen storage, *Chem. Rev.* 107 (10) (2007) 4111–4132.
- Morten B. Ley, Lars H. Jepsen, Young-Su Lee, Young Whan Cho, José M. Bellosta Von Colbe, Martin Dornheim, Masoud Rokni, Jens Oluf Jensen, Mikael Sloth, Yaroslav Filinchuk, Complex hydrides for hydrogen storage—new perspectives, *Mater. Today* 17 (3) (2014) 122–128.
- R.S. Michalski, J.G. Carbonell, T.M. Mitchell, *Machine Learning: An Artificial Intelligence Approach*, Springer Science & Business Media, 2013.
- Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
- S. Papanikolaou, Learning local, quenched disorder in plasticity and other crackling noise phenomena, *npj Comput. Mater.* 4 (1) (2018) 27.
- S. Ramakrishna, T.-Y. Zhang, W.-C. Lu, Q. Qian, J.S.C. Low, J.H.R. Yune, D.Z.L. Tan, S. Bressan, S. Sanvito, S.R. Kalidindi, *Materials informatics, J. Intell. Manuf.* 30 (2019) 2307–2326.
- R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: Recent applications and prospects, *npj Comput. Mater.* 3 (1) (2017) 54.
- K. Takahashi, Y. Tanaka, *Materials informatics: a journey towards material design and synthesis*, *Dalton Trans.* 45 (26) (2016) 10497–10499.
- L. Ward, C. Wolverton, Atomistic calculations and materials informatics: A review, *Curr. Opin. Solid State Mater. Sci.* 21 (3) (2017) 167–176.
- K. Rajan, *Materials informatics, Mater. Today* 8 (10) (2005) 38–45.
- J. Greeley, T.F. Jaramillo, J. Bonde, I.B. Chorkendorff, J.K. Nørskov, Computational high-throughput screening of electrocatalytic materials for hydrogen evolution, *Nature Mater.* 5 (11) (2006) 909–913.
- W.T. Hong, R.E. Welsch, Y. Shao-Horn, Descriptors of oxygen-evolution activity for oxides: a statistical evaluation, *J. Phys. Chem. C* 120 (1) (2016) 78–86.
- C.E. Wilmer, M. Leaf, C.Y. Lee, O.K. Farha, B.G. Hauser, J.T. Hupp, R.Q. Snurr, Large-scale screening of hypothetical metal-organic frameworks, *Nature Chem.* 4 (2) (2012) 83–89.
- L.-C. Lin, A.H. Berger, R.L. Martin, J. Kim, J.A. Swisher, K. Jariwala, C.H. Rycroft, A.S. Bhowm, M.W. Deem, M. Haranczyk, et al., In silico screening of carbon-capture materials, *Nature Mater.* 11 (7) (2012) 633–641.
- M. Fernandez, N.R. Trefiak, T.K. Woo, Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity, *J. Phys. Chem. C* 117 (27) (2013) 14095–14105.
- M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Phys. Rev. Lett.* 108 (5) (2012) 058301.
- G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O.A. Von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space, *New J. Phys.* 15 (9) (2013) 095003.
- W. Kohn, A.D. Becke, R.G. Parr, Density functional theory of electronic structure, *J. Phys. Chem.* 100 (31) (1996) 12974–12980.
- A.J. Cohen, P. Mori-Sánchez, W. Yang, Challenges for density functional theory, *Chem. Rev.* 112 (1) (2012) 289–320.
- J.K. Nørskov, F. Abild-Pedersen, F. Studt, T. Bligaard, Density functional theory in surface chemistry and catalysis, *Proc. Natl. Acad. Sci.* 108 (3) (2011) 937–943.
- R.G. Parr, W. Yang, Density-functional theory of the electronic structure of molecules, *Annu. Rev. Phys. Chem.* 46 (1) (1995) 701–728.
- J.W.D. Connolly, A.R. Williams, Density-functional theory applied to phase transformations in transition-metal alloys, *Phys. Rev. B* 27 (8) (1983) 5169.
- D. Kwak, M. Wang, K.J. Koski, L. Zhang, H. Sokol, R. Maric, Y. Lei, Molybdenum trioxide (α -MoO₃) nanoribbons for ultrasensitive ammonia (NH₃) gas detection: integrated experimental and density functional theory simulation studies, *ACS Appl. Mater. Interfaces* 11 (11) (2019) 10697–10706.
- Q. He, B. Yu, Z. Li, Y. Zhao, Density functional theory for battery materials, *Energy Environ. Mater.* 2 (4) (2019) 264–279.
- A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization, *Phys. Rev. Lett.* 115 (20) (2015) 205901.
- D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, *Nature Commun.* 7 (2016) 11241.
- O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, *Nature Commun.* 8 (1) (2017) 15679.
- M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Phys. Rev. Lett.* 108 (5) (2012) 058301.
- J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, *J. Chem. Phys.* 134 (7) (2011).
- F. Pietrucci, W. Andreoni, Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale, *Phys. Rev. Lett.* 107 (8) (2011) 085504.
- S. De, A.P. Bartók, G. Csányi, M. Ceriotti, Comparing molecules and solids across structural and alchemical space, *Phys. Chem. Chem. Phys.* 18 (20) (2016) 13754–13769.
- A. Sadeghi, S.A. Ghasemi, B. Schaefer, S. Mohr, M.A. Lill, S. Goedecker, Metrics for measuring distances in configuration spaces, *J. Chem. Phys.* 139 (18) (2013).
- Stefano Curtarolo, Gus L.W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, Ohad Levy, The high-throughput highway to computational materials design, *Nature Mater.* 12 (3) (2013) 191–201.

- [43] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, Christopher Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), *Jom* 65 (2013) 1501–1509.
- [44] *differential*, is here addressing charge density minus superposition of atomic densities.
- [45] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L. Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, *J. Phys. Condens. Matter* 21 (39) (2009) 395502.
- [46] Paolo Giannozzi, Oliviero Andreussi, Thomas Brumme, Oana Bunau, M. Buongiorno Nardelli, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Matteo Cococcioni, et al., Advanced capabilities for materials modelling with quantum ESPRESSO, *J. Phys. Condens. Matter* 29 (46) (2017) 465901.
- [47] Paolo Giannozzi, Oscar Baseggio, Pietro Bonfà, Davide Brunato, Roberto Car, Ivan Carnimeo, Carlo Cavazzoni, Stefano De Gironcoli, Pietro Delugas, Fabrizio Ferrari Ruffino, et al., Quantum ESPRESSO toward the exascale, *J. Chem. Phys.* 152 (15) (2020).
- [48] John P. Perdew, Kieron Burke, Matthias Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* 77 (18) (1996) 3865.
- [49] Riccardo Sabatini, Tommaso Gorni, Stefano De Gironcoli, Nonlocal van der Waals density functional made simple and efficient, *Phys. Rev. B* 87 (4) (2013) 041108.
- [50] M.P.A.T. Methfessel, A.T. Paxton, High-precision sampling for Brillouin-zone integration in metals, *Phys. Rev. B* 40 (6) (1989) 3616.
- [51] https://dalcorso.github.io/thermo_pw/.
- [52] D. Ikuta, E. Ohtani, A. Sano-Furukawa, Y. Shibazaki, H. Terasaki, L. Yuan, T. Hattori, Interstitial hydrogen atoms in face-centered cubic iron in the Earth's core, *Sci. Rep.* 9 (2019) 7108.
- [53] W. Tang, E. Sanville, G. Henkelman, A grid-based bader analysis algorithm without lattice bias, *J. Phys.: Condens. Matter* 21 (8) (2009) 084204.
- [54] Here we call *Bader radii* the minimum distances between each atom and the boundary of its Bader volume.
- [55] M. Łukasz, Mendeleev – A Python resource for properties of chemical elements, ions and isotopes, 2014, <https://github.com/lmmmentel/mendeleev>, 0.11.0.
- [56] <https://github.com/materialsproject/pyrho>.
- [57] M.S. Shell, Thermodynamics and Statistical Mechanics: An Integrated Approach, Cambridge University Press, 2015.
- [58] Q. Li, X. Lin, Q. Luo, Y. Chen, J. Wang, B. Jiang, F. Pan, Kinetics of the hydrogen absorption and desorption processes of hydrogen storage alloys: A review, *Int. J. Miner. Metall. Mater.* 29 (2022) 32–48.
- [59] K. Koizumi, K. Nobusada, M. Boero, Hydrogen storage mechanism and diffusion in metal–organic frameworks, *Phys. Chem. Chem. Phys.* 21 (15) (2019) 7756–7764.
- [60] Z. Chen, Z. Ma, J. Zheng, X. Li, E. Akiba, H. Li, Perspectives and challenges of hydrogen storage in solid-state hydrides, *Chin. J. Chem. Eng.* 29 (2021) 1–12.
- [61] J. Čermák, L. Král, Hydrogen diffusion in Mg–H and Mg–Ni–H alloys, *Acta Mater.* 56 (12) (2008) 2677–2686.
- [62] K. Pearson, LIII, On lines and planes of closest fit to systems of points in space, *Lond. Edinb. Dublin Philos. Mag. J. Sci.* 2 (11) (1901) 559–572.
- [63] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.* 24 (6) (1933) 417.
- [64] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, *Philos. Trans. R. Soc. a Math. Phys. Eng. Sci.* 374 (2065) (2016) 20150202.
- [65] S. Lloyd, Least squares quantization in PCM, *IEEE Trans. Inf. Theory* 28 (2) (1982) 129–137.
- [66] P. Berkhin, Survey of Clustering Data Mining Techniques, Technical Report, Accrue Software San Jose CA, 2002.
- [67] D. Arthur, S. Vassilvitskii, K-means++ the advantages of careful seeding, in: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 1027–1035.
- [68] R.L. Thorndike, Who belongs in the family? *Psychometrika* 18 (4) (1953) 267–276.
- [69] it is to notice how this quantity is related to compactness of the clusters.
- [70] C. Li, J.-B. Baek, Recent advances in noble metal (Pt, Ru, and Ir)-based electrocatalysts for efficient hydrogen evolution reaction, *ACS Omega* 5 (1) (2020) 31–40.
- [71] F. Marques, M. Balcerzak, F. Winkelmann, G. Zepon, M. Felderhoff, Review and outlook on high-entropy alloys for hydrogen storage, *Energy Environ. Sci.* 14 (10) (2021) 5191–5227.
- [72] it will be shown explicitly soon in the same section.
- [73] The radius of the M charge here behaves differently from the volume, and it may be due to its definition in the bader extraction program, as the minimum distance from the Bader surface. In this view, volume can be seen as a more reliable estimate of charge extensions due to anisotropies and bondings. We believe this to be reason also for the very high correlation between V(M) and V(H) and less pronounced one between R(M) and R(H).
- [74] These are properties of the metal bulk crystals before being defected with interstitial hydrogen.
- [75] S. Satyapal, J. Petrovic, C. Read, G. Thomas, G. Ordaz, The U.S. Department of Energy's National Hydrogen Storage Project: Progress towards meeting hydrogen-powered vehicle requirements, *Catalysis Today* 120 (3) (2007) 246–256.
- [76] G. Lu, D. Orlikowski, I. Park, O. Politano, E. Kaxiras, Energetics of hydrogen impurities in aluminum and their effect on mechanical properties, *Phys. Rev. B* 65 (6) (2002) 064102.
- [77] L. Messina, M. Nastar, N. Sandberg, Pär Olsson, Systematic electronic-structure investigation of substitutional impurity diffusion and flux coupling in BCC iron, *Phys. Rev. B* 93 (18) (2016) 184302.