Charge-density descriptors from a CHGCAR

What the 56 pydemi descriptors measure, and how they are computed

Shubham Kumar Maurya, CMS Lab, IIT Kanpur • 13 October 2026

In one paragraph. Every descriptor below is a single number computed from one file, the VASP CHGCAR: the crystal structure in its header and the electron density ρ on a periodic grid. The descriptors fall into three groups. *Bonding* descriptors (19) ask how the charge is arranged around each atom and in the bonds. *Structural* descriptors (19) look at the density of the whole cell: how well it connects, its peaks and valleys, its directionality and its information content. *Heterogeneity* descriptors (18) ask how different the atoms of one structure are from each other. Every value is intensive (unchanged when the cell is replaced by a supercell) and unchanged by translating or rotating the crystal.

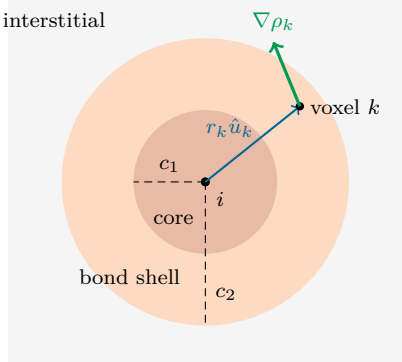
1 Notation and the common machinery

Grid. The cell (volume V) is divided into N_v voxels k of volume $\Delta V = V/N_v$. ρ_k is the density at voxel k in e/Å^3 (the CHGCAR value divided by V). The total charge is $Q_{\text{tot}} = \sum_k \rho_k \Delta V$, and $\langle\rho\rangle = Q_{\text{tot}}/V$ is the mean density.

Nearest nucleus. For each voxel, $i(k)$ is the nearest atom (periodic images included), r_k its distance and $\hat{u}_k$ the unit vector from that nucleus to the voxel.

Shells. Each voxel falls in one of three shells around its nearest nucleus (defaults $c_1 = 0.8$ Å, $c_2 = 1.5$ Å; optionally multiples of the covalent radius):

core:	$r_k \le c_1$
bond shell:	$c_1 < r_k \le c_2$
interstitial:	$r_k > c_2$



Each voxel k is measured from its nearest nucleus.

Derivatives. $\nabla\rho$, $\nabla^2\rho$ and the Hessian eigenvalues $\lambda_1 \le \lambda_2 \le \lambda_3$ are computed spectrally (FFT) on the periodic grid. Quantities defined in atomic units (NCI, information measures) are converted to bohr first.

Degenerate cases. A descriptor that is undefined for a special input (e.g. a perfectly uniform density, or an empty shell) returns a documented constant, and a flag is set in the metadata; it never returns a bare NaN.

Numerical stability. Descriptors whose value changed by more than about 10% on the dataset when the derivatives were taken by finite differences instead of FFT, or when the grid was coarsened to 80%, were dropped (see the end).

2 Bonding descriptors (19)

2.1 Where the charge sits: radial moments and shell fractions

The simplest picture of a solid is a set of atoms. These descriptors ask how far, on average, the charge lies from its nearest nucleus, and how it is shared between the core, the bond shell and the space in between.

`m1` [Å] $m_1 = \dfrac{\sum_k \rho_k r_k}{\sum_k \rho_k}$

Mean distance of the charge from its nearest nucleus. Small for compact, core-dominated atoms; large for diffuse valence charge.

`m2` [Å^2] $m_2 = \dfrac{\sum_k \rho_k r_k^2}{\sum_k \rho_k}$

Second radial moment; weights the outer charge more strongly.

`sigma_r2` [Å^2] $\sigma_r^2 = m_2 - m_1^2$

Radial spread of the charge about its mean distance.

`f_core, f_bond, f_int` $f_{\text{shell}} = \dfrac{\sum_{k\in\text{shell}} \rho_k}{\sum_k \rho_k}, \quad f_{\text{core}} + f_{\text{bond}} + f_{\text{int}} = 1$

Fraction of the charge in the core, bond and interstitial shells. A large f_{int} points to delocalized (metallic) charge between the atoms.

2.2 Is the charge spherical? Gradient anisotropy

Around an isolated, spherical atom the density gradient points straight at the nucleus. Bonds, lone pairs and other distortions tilt it sideways. ζ measures the share of the gradient that is *not* radial.

`zeta` [range $[0, 1]$] $\zeta = 1 - \dfrac{\sum_k |\nabla\rho_k \cdot \hat{u}_k|}{\sum_k |\nabla\rho_k|}$

0 for spherical, non-overlapping atoms; grows as the charge is pulled off the radial direction (directional bonding).

2.3 Where charge is concentrated: the Laplacian

Where $\nabla^2\rho < 0$ the density is locally concentrated (higher than its surroundings), as in atomic shells, covalent bonds and lone pairs; where $\nabla^2\rho > 0$ it is depleted.

`lnf` $\text{lnf} = \dfrac{1}{N_v} \sum_k \mathbf{1}[\nabla^2\rho_k < 0]$

Volume fraction of the cell where charge is concentrated.

`lnf_charge_weighted` $\dfrac{\sum_{\nabla^2\rho_k<0} \rho_k}{\sum_k \rho_k}$

Fraction of the *charge* in concentration regions. Less sensitive to the grid than `lnf`, because low-density voxels carry little weight.

`lap_concentration` $\dfrac{\sum_{\nabla^2\rho_k<0} |\nabla^2\rho_k|}{\sum_k |\nabla^2\rho_k|}$

On a periodic grid $\int \nabla^2\rho \, dV = 0$ exactly, so this is always 1/2. It carries no information and is kept only for completeness.

`lap_concentration_valence` same ratio, restricted to $r_k > c_1$

The informative version: the balance of concentration and depletion in the valence region, outside the cores.

2.4 Shape of the bonds: ellipticity

At a bond, the two Hessian eigenvalues across the bond are negative. If they are equal the bond is cylindrical (a σ bond); if they differ the charge is flattened, as in π character. The usual ellipticity $\varepsilon = \lambda_1/\lambda_2 - 1$ diverges as $\lambda_2 \to 0$, so pydemi uses its bounded form. Let $B = \{k \in \text{bond shell} : \lambda_2 < 0\}$.

`ellip_bond_bounded_avg` [range $[0, 1)$] mean over B of $1 - \dfrac{\lambda_2}{\lambda_1} = \dfrac{\varepsilon}{1+\varepsilon}$

0 for a cylindrical bond; approaches 1 for strongly flattened charge. Each voxel's value is at most 1, so a few voxels with $\lambda_2 \approx 0$ cannot dominate the mean.

2.5 Weak interactions (NCI)

Non-covalent interactions show up where the density is low and nearly flat, which a small reduced gradient s picks out. The sign of λ_2 separates attractive ($\lambda_2 < 0$) from repulsive contacts. The NCI region is $\{k : s_k < 0.5,\ \rho_k < 0.05$ a.u.$\}$, where $s = |\nabla\rho| \,/\, (2(3\pi^2)^{1/3}\rho^{4/3})$.

`f_NCI` volume fraction of the NCI region
How much of the cell holds weak-interaction density.

`NCI_attractive` fraction of NCI voxels with $\lambda_2 < 0$
Share of the weak interactions that are attractive.

`sign_lambda2_rho_mean` [e/bohr^3] mean over the NCI region of $\mathrm{sign}(\lambda_2)\,\rho$
Signed strength of the weak interactions: negative means mostly attractive.

2.6 Electrostatics (Hartree potential)

The Hartree potential energy of an electron is obtained from ρ by one FFT, $V(\mathbf{G}) = 4\pi k_e\,\rho(\mathbf{G})/|\mathbf{G}|^2$, with $k_e = 14.40\,\mathrm{eV\,\AA}$. The cell average is set to zero.

`V_spread` [eV] $\mathrm{std}_i\, V(\mathbf{R}_i)$ over the atomic sites
How different the electrostatic environments of the atoms are; a rough signature of charge transfer.

`V_int_min` [eV] $\min_{k\in\mathrm{int}} V_k$
Lowest potential in the interstitial region.

2.7 Density in the bonds

Neighbours: j is a first-shell neighbour of i if $|\mathbf{R}_{ij}| \le 1.1\,d_i$, where d_i is the distance from i to its nearest neighbour. ρ at each bond midpoint is obtained by exact Fourier interpolation.

`rho_mid_mean` [e/Å^3] mean of ρ at the nearest-neighbour bond midpoints
How much charge sits in the bonds: high for covalent bonds, low for ionic contacts.

`rho_mid_std` [e/Å^3] standard deviation of the same values
How uneven the bonds are.

3 Structural descriptors (19)

3.1 Connectivity: percolation

Lower a threshold c from the density maximum. At first the region $\{\rho > c\}$ consists of isolated islands around the atoms. At some level the islands join into a path that runs through the whole crystal along a lattice direction $\mathbf{a}_\alpha$ (it winds around the periodic cell).

`rho_perc_a, rho_perc_b, rho_perc_c` [e/Å^3] $\rho_{\mathrm{perc},\alpha} = \max\{c :\ \{\rho > c\}$ spans along $\mathbf{a}_\alpha\}$
Density at the bottleneck of the best connecting path. High in metals, close to zero in ionic and molecular solids.

`perc_anisotropy` $(\max_\alpha - \min_\alpha)\,\rho_{\mathrm{perc},\alpha} \,/\, \mathrm{mean}_\alpha\,\rho_{\mathrm{perc},\alpha}$
Whether the charge connects equally well in all three directions (0), or better along some (layered or chain-like structures).

3.2 Peaks and valleys: the critical-point census

Counts are divided by the cell volume, so they are per Å^3.

`n_max, n_min` [1/Å^3] voxels higher (lower) than all 26 neighbours, divided by V
Local maxima (typically at or near the nuclei) and minima (centres of the voids).

`n_NNM` [1/Å^3] maxima farther than 0.8 Å from every nucleus, divided by V

Non-nuclear maxima: charge pooled away from the atoms, e.g. in electrides or some metals.

`Q_NNM` (charge in the steepest-ascent basins of those maxima)/Q_{tot}

How much charge the non-nuclear maxima hold.

3.3 The density floor

`rho_min` [e/Å^3] $\min_k \rho_k$

Lowest density in the cell. In a CHGCAR (a PAW pseudo-density) this is often slightly negative and near a nucleus, not in the voids.

`rho_min_ratio` $\rho_{\min}/\langle\rho\rangle$

The same, relative to the mean density.

`rho_int_mean` [e/Å^3] mean of ρ over the interstitial shell

How much charge fills the space between the atoms.

3.4 Directionality: the charge-anisotropy tensor

Averaging the gradient's directions over the whole cell gives a 3×3 tensor with trace 1:

$$T_{ab} = \frac{\sum_k \partial_a\rho_k\,\partial_b\rho_k}{\sum_k |\nabla\rho_k|^2}, \qquad \text{eigenvalues } t_1 \le t_2 \le t_3, \quad t_1+t_2+t_3 = 1.$$

`T_eigenvalues_t1, _t2, _t3` the sorted eigenvalues of T

All 1/3 when the gradients point equally in all directions (e.g. any cubic crystal). Unequal values show that the density varies more along some directions, as in layered or chain structures.

`charge_FA` [range $[0,1]$] $\text{FA} = \sqrt{\frac{3}{2}}\,\frac{\|T - \frac{1}{3}I\|_F}{\|T\|_F}$

Fractional anisotropy: one number for how far T is from isotropic. 0 when isotropic; 1 when every gradient is parallel.

3.5 Information content

These treat the normalized density $\tilde\rho = \rho/N_e$ (with $N_e = \int \rho\,dV$) as a probability distribution over the cell. Each is normalized by the cell volume, which makes it intensive.

`shannon_entropy` [range ≤ 0] $S = -\int \tilde\rho \ln\tilde\rho\,dV - \ln V$

How spread out the charge is: 0 for a uniform density, more negative as the charge concentrates on the atoms.

`disequilibrium` [range ≥ 1] $D = V\int \tilde\rho^{\,2}\,dV$

How far the charge is from uniform: 1 for a uniform density, larger when it is concentrated.

`LMC_complexity` $C = D\,e^{S}$

López-Ruiz–Mancini–Calbet complexity. It balances order against disorder and equals 1 for a uniform density.

`fisher_information` [1/bohr^2] $I = \int \frac{|\nabla\tilde\rho|^2}{\tilde\rho}\,dV$

Sharpness of the density: large when it changes steeply (compact cores, strong localization).

4 Heterogeneity descriptors (18)

The descriptors above describe the whole cell. These ask *how different the atoms are*. Each voxel is assigned to an atom (default: its nearest nucleus, i.e. Voronoi cells; smooth partitions give fractional

weights). Then three per-atom quantities are computed, each the per-atom version of a bonding descriptor ($w_i(k) = 1$ if voxel k belongs to atom i):

$$m_1^{(i)} = \frac{\sum_k w_i(k)\rho_k r_{ik}}{\sum_k w_i(k)\rho_k}, \qquad f_{\text{bond}}^{(i)} = \frac{\sum_{c_1 < r_{ik} \le c_2} w_i(k)\rho_k}{\sum_k w_i(k)\rho_k}, \qquad \zeta^{(i)} = 1 - \frac{\sum_k w_i(k)|\nabla\rho_k \cdot \hat{u}_{ik}|}{\sum_k w_i(k)|\nabla\rho_k|}.$$

For each $X \in \{m_1, f_{\text{bond}}, \zeta\}$, six statistics over the atoms are reported (18 in total). Let x_e be the fraction of sites of element e, $\bar{X}_e$ the mean over the sites of element e, and $\bar{X}$ the mean over all sites.

descriptor	definition	meaning
`X_site_std`	$\text{std}_i X^{(i)}$	overall spread between atoms
`X_site_range`	$\max_i X^{(i)} - \min_i X^{(i)}$	the most extreme difference
`X_site_max`, `X_site_min`	$\max_i X^{(i)}$, $\min_i X^{(i)}$	the extreme atoms
`X_within_element_var`	$\sum_e x_e \,\text{Var}_{i\in e} X^{(i)}$	atoms of the *same* element in different environments
`X_between_element_var`	$\sum_e x_e\,(\bar{X}_e - \bar{X})^2$	differences *between* elements

The last two split the variance exactly, $\text{Var}_i X$ = within + between (as in one-way ANOVA). The within-element variance is undefined (NaN, flagged) when no element has two or more sites. The between-element variance is 0 (flagged) for a single-element crystal. Units are those of X (Å for m_1, squared for the variances; f_{bond} and ζ are dimensionless).

Summary

group	n	what it captures
moments and shells	6	where the charge sits relative to the nuclei
gradient anisotropy	1	how non-spherical the atoms are
Laplacian	4	where charge is concentrated or depleted
ellipticity	1	cylindrical vs. flattened bonds
NCI	3	weak, non-covalent interactions
Hartree potential	2	electrostatic environment of the atoms
bond midpoints	2	charge in the bonds
percolation	4	how well the charge connects through the crystal
critical points	4	peaks, valleys and non-nuclear maxima
density floor	3	charge in the voids
anisotropy tensor	4	directionality of the whole density
information measures	4	spread, concentration and sharpness
heterogeneity	18	differences between atoms, within and between elements
total	**56**	

Defaults: $c_1 = 0.8$ Å, $c_2 = 1.5$ Å; FFT derivatives; nearest-atom partition; `featurize` returns all 56.

Not included, and why.

- *ELF and local energy densities* (g, v, H): a CHGCAR has no orbitals, so both would have to be estimated from the PAW pseudo-density with a density-only approximation of the kinetic energy. The results are not the true quantities.
- *Mean and spread of the unbounded ellipticity* ε: they diverge as $\lambda_2 \to 0$. On the dataset the mean changed by a median 22% (FD4 vs. FFT derivatives) and the spread by 67%; the bounded form above changed by 0.2%.
- *Saddle-point counts*: they changed by a median 17–20% when the grid was coarsened to 80%.