BEYOND ENERGIES AND FORCES: OMOL-DESCRIPTORS-4M, AN OPEN POST-DFT PIPELINE AND EVALUATION PROTOCOL FOR MACHINE LEARNING IN CHEMISTRY

Anonymous authors
Paper under double-blind review

ABSTRACT

We present `qtaim_generator`, an end-to-end computational chemistry post-analysis framework that processes repositories of molecular DFT wavefunctions and yields six partial-charge schemes (Hirshfeld, CM5, ADCH, Becke, Mulliken, Löwdin), four bond-order schemes, full QTAIM topology, fuzzy descriptors, and ORCA-derived globals. We built this by wrapping per-tier analysis routines around Multiwfn and ORCA's `orca_2mkl` utility, packaging the result into a scalable workflow for HPC clusters. Users can select compute tiers to trade fidelity for cost, then compact the resulting calculations into portable sharded Lightning Memory-Mapped Databases (LMDBs) keyed by job identity for downstream analysis and machine learning. To demonstrate the pipeline, we processed the recently-released OMol25 electronic structure dataset: a 4M-structure public subset of the 110M+-calculation 2025 release that ships the complete DFT outputs, including density matrices, energy gradients, ORCA `.out` files, and `.gbw` wavefunctions. By post-processing the wavefunctions, we extract atom-, bond-, and molecule-level descriptors across $\sim$4M structures spanning 34 chemical verticals, supplementing the original dataset's energy and force labels with physically-interpretable quantities. We name this dataset OMol-Descriptors-4M and present it as the first multi-vertical descriptor dataset at this scale beyond energies and forces. Analogous to the source dataset, we establish an evaluation protocol with explicit evaluative claims and scope boundaries: composition-ordered train/val/test splits via a hashed molecular composition and five domain-specific held-out stress tests (metal-ligand bond pairs, lanthanide-ligand bond pairs, reactivity, large systems, and high-charge regimes; $\sim$60k structures total). Together with the open generation pipeline, these components let researchers reuse the protocol on new wavefunction repositories across molecular chemistry.

1 INTRODUCTION

Machine learning and deep learning have upended the relationship between large datasets and how we interpret, predict, and generate information. Prior to this revolution, humans relied on low-dimensional embeddings and descriptors to parse and understand large swaths of data. Today, a wide spectrum of models operate on language (Brown et al., 2020), vision (Redmon et al., 2016; Kirillov et al., 2023), and biological (Jumper et al., 2021) data. In the chemical sciences, we have seen the decades long development of descriptors and models heat up in the last 10 years. Cheminformatics via fingerprinting descriptors (Rogers & Hahn, 2010) and Morgan (Morgan, 1965) coupling to tree-based supervised-learning algorithms (Ahneman et al., 2018) have given way to advanced architectures and representations rooted in graph-based embeddings of chemical structures. These newer approaches include continuous-filter graph neural networks (Schütt et al., 2018), directional message-passing architectures (Gasteiger et al., 2020), and equivariant foundation models capable of stable atomistic simulation across nearly the entire periodic table (Batatia et al., 2025). High-profile applications include generative inorganic materials design with MatterGen (Zeni et al., 2025), the OMat24 dataset and models for materials neural network potentials (Barroso-Luque et al., 2024), and 3D pre-training on hundreds of millions of conformers for broad molecular property prediction (Lu et al., 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Naturally, chemistry and material science, with their mature methods for generating simulated data via Density Functional Theory (DFT) have led to the creation of many datasets. QM9, for example, is one of the most popular datasets and covers a set of ∼134K organic molecules with up to 9 heavy atoms (Ramakrishnan et al., 2014). Beyond this, SPICE (Eastman et al., 2023) and ANI-1x (Smith et al., 2020) extended this coverage to biomolecular and reactive domains. OC20 (Chanussot et al., 2021), Materials Project (Horton et al., 2025) and tmQM+ (Gee et al., 2025) pushed the boundary towards inorganic materials and transition-metal complexes. Alongside datasets, benchmarking suites such as MoleculeNet (Wu et al., 2018) and MOSES (Polykovskiy et al., 2020) have standardized evaluation across property prediction and molecular generation tasks. The recent OMol25 (Levine et al., 2026) is notable among molecular MLIP datasets due to its size, diversity, and high-level of accuracy. It pushes beyond 100M structures, 80+ elements, and many chemical verticals to provide the community with a rich repository for training neural network potentials. To date, the team has shared the wavefunction, density matrices, and orca.out files of 4M of those datapoints. This additional information serves as the basis for this work where we post-processed OMol25 wavefunctions to yield a new dataset of descriptors. We name the result **OMol-Descriptors-4M** and contribute: (i) the dataset itself, (ii) the open generation pipeline, and (iii) an evaluation protocol with composition-ordered splits, five held-out stress-test suites, and cross-scheme noise floors (the median absolute residual between schemes on identical atoms, App. K).

Intended uses. Four uses motivate the release. First, charge-aware MLIPs such as BAMBOO (Gong et al., 2025) and AIMNet2 (Anstine et al., 2025) carry atomic charges for long-range electrostatics, and `energy.lmdb` places the OMol25 energies and forces on the same keys as six charge schemes, so charge and energy heads can be trained jointly. Second, polarizable force fields such as AMOEBA (Ponder et al., 2010) derive their electrostatic parameters from quantum-mechanical densities, and a charge surrogate trained here supplies density-derived charges without per-molecule DFT. Third, QTAIM descriptors improve machine learning on transition-metal complexes (Gee et al., 2025), and this corpus extends such descriptors to 3d–5d metals, lanthanides and reactive geometries. Fourth, quantum-mechanical descriptors improve graph neural networks as input features (Li et al., 2024; Vargas et al., 2024) and as auxiliary information for reaction-barrier prediction (Vargas et al., 2021), and using them at scale requires predicting them rather than computing them.

2 Related Work

Descriptor-rich quantum-chemistry datasets divide along two axes: the chemical envelope they cover and the descriptor families they incorporate. Smaller, single-purpose releases such as QM9, OE62, BDE-db, and tmQM+ (Ramakrishnan et al., 2014; Stuke et al., 2020; St. John et al., 2020; Gee et al., 2025) sit below 10^6 structures and typically contain only one descriptor family per dataset. QMugs (Isert et al., 2022) is the closest multi-descriptor precedent at scale: 665,911 drug-like molecules (1,992,984 conformers) with Mulliken and Löwdin charges and Mayer and Wiberg–Löwdin bond orders at ωB97X-D/def2-SVP, but no real-space partitioning or QTAIM topology, and ten main-group elements. AIMEl (Meza-González et al., 2024) provides QTAIM atomic basin properties for 44,470 CHNO molecules from QM9. In contrast, modern force-field training corpora such as ANI-1x/2x, SPICE, GEOM, transition1x, and OMol25 itself (Smith et al., 2020; Devereux et al., 2020; Eastman et al., 2023; Axelrod & Gómez-Bombarelli, 2022; Schreiner et al., 2022; Levine et al., 2026) routinely cross 10^6–10^8 structures but focus only on energies, forces, and occasionally, a limited set of other features. Even when the latter contain information such as dipoles, orbital information, and/or partial charges, they typically use only one or two cruder schemes. OMol-Descriptors-4M encompasses a wider diversity and size vs. existing descriptor datasets, all while providing far more descriptors as compared to existing datasets for training MLIPs. As this is a broader descriptor dataset, we will only use comparators with a wider set of features beyond one partial charge scheme that are also not contained within the larger OMol25 dataset.

2.1 Comparable-scale descriptor comparators

As mentioned, any comparison of descriptors between our dataset and others is moot as existing datasets simply do not cover the diversity of features within OMol-Descriptors-4M, and typically, do not contain the chemical diversity of OMol25. In lieu of this, we opted to compare the structural diversity within our dataset to others of comparable descriptor richness, namely, QM7-X, SchNet4AIM,

and tmQM+. QM7-X (Hoja et al., 2021) reaches 4.2M conformers with a rich per-atom data inventory: Hirshfeld charges, dipoles, atomic C_6 coefficients, polarizabilities, van der Waals radii, and a PBE0+MBD energy decomposition – but only 6,950 unique compositions across CHNOSCl compositions with at most seven heavy atoms. SchNet4AIM (Gallegos et al., 2024), though smaller with only $\sim$3.9k structures, is a closer analog on the axis of dataset diversity. It contains only CHNO atoms but incorporates QTAIM charges, atomic localization indices (λ), and pairwise delocalization indices (δ). The tmQM+ dataset (Gee et al., 2025) is a near precedent on the descriptor-richness axis for transition-metal chemistry: it incorporates $\sim$60,000 transition-metal complexes with full QTAIM topology (critical-point scalars, Lagrangian and Hamiltonian kinetic energy densities, electron localization function, ESP components, gradient and Laplacian norms, ellipticity, and pairwise QTAIM bond indices) computed at three levels of theory. Earlier, Vargas et al. (2024) showed that QTAIM descriptors improve geometric deep learning of molecular and reaction properties.

Table 1: Descriptor-rich comparator datasets. "Compositions" counts unique molecular formulas. LOT: level of theory. QMugs compositions (175,414) are computed from its released SMILES.

Dataset	Structures	Compositions	Chemical Vertical	Descriptors
QM7-X	4.2M	6.9k	CHNOSCl, ≤ 7 heavy atoms	Hirshfeld; C_6, α, vdW; PBE0+MBD
SchNet4AIM	3.9k	3.9k	CHON	QTAIM Q, λ, δ
QMugs	2.0M	175k	Drug-like (ChEMBL), 10 main-group elements	Mulliken, Löwdin; Mayer, Wiberg–Löwdin bond orders; ωB97X-D/def2-SVP
tmQM+ (Gee et al., 2025)	60k	60k	3d/4d/5d transition-metal complexes	Full QTAIM topology, ELF, ESP, kinetic energy densities; 3 LOTs
Ours	**4.0M**	**1.30M**	TM, charged, reactive, protein-derived	6 charge, 4 bond-order schemes, full QTAIM, fuzzy, ORCA globals

We confirm that the descriptor breadth above corresponds to genuine chemistry-space coverage by projecting OMol-Descriptors-4M against the organic comparators in SOAP-UMAP space, fit on a balanced 500-structure-per-source anchor with a 12-element SOAP species set (Fig. 12, App. N). A transition-metal-inclusive embedding against tmQM+ is in the same appendix; in both, OMol-Descriptors-4M appears to cover a superset of the comparator structures.

3 Dataset

3.1 Descriptor Families

Partial charges. Six independent atomic partial-charge schemes are computed per structure. Hirshfeld (Hirshfeld, 1977), CM5 (Marenich et al., 2012), ADCH (Lu & Chen, 2012a), and Becke (Becke, 1988) charges are produced by Multiwfn (Lu & Chen, 2012b) from the DFT electron density; Mulliken and Löwdin charges are extracted directly from the ORCA output (Neese et al., 2022). Hirshfeld charges partition the electron density by the stockholder principle: each atom claims a share of the density at every point proportional to its free-atom promolecular density, yielding compact, chemically intuitive atomic populations. CM5 applies a pairwise connectivity-based correction to Hirshfeld charges to reproduce experimental dipole moments across diverse environments (Marenich et al., 2012). Becke charges integrate the molecular density over smooth, fuzzy atomic basins defined by Becke's numerical weight functions (Becke, 1988), providing a partition that is independent of the basis set and free of the basis-function artifacts that affect Mulliken and Löwdin populations. ADCH further corrects Hirshfeld charges for atomic dipole contributions to reduce basis-set dependence (Lu & Chen, 2012a). Multiwfn applies the same atomic-dipole correction to the Becke populations, so the released `becke` scheme is ADC-corrected Becke rather than raw Becke populations; ADC-Becke and ADCH share the correction but not the underlying partition. Because ORCA and Multiwfn are independent codes, comparisons between the ORCA-derived (Mulliken, Löwdin) and Multiwfn-derived schemes (App. J) are cross-code as well as cross-scheme.

Bond orders. Four schemes are universal across the corpus: Mayer (Mayer, 1983) and Löwdin bond orders from ORCA, fuzzy bond orders from Multiwfn (Lu & Chen, 2012b; Becke, 1988), and QTAIM bond presence inferred from bond critical points. Mayer bond orders derive from the density matrix and capture bond multiplicity beyond formal valence, including partial and dative character. Löwdin bond orders use a symmetrically orthogonalized basis, reducing the basis-set sensitivity that affects Mulliken-derived indices. Fuzzy bond orders integrate the exchange-correlation hole over Becke

fuzzy-atom basins, providing a real-space, basis-set-independent measure of covalent bond character. QTAIM bond presence is a binary signal at each bond critical point; the corresponding density and Laplacian scalars are stored under QTAIM topology. The schemes report different atom-pair sets: ORCA prints Mayer bond orders above 0.10 and Löwdin bond orders above 0.05, Multiwfn lists pairs whose total fuzzy bond order is at least 0.05, and a QTAIM bond exists only where a $(3,-1)$ critical point is located.

QTAIM topology. Critical points of the electron density are located via Multiwfn following Bader's Quantum Theory of Atoms in Molecules (Bader, 1990). Each bond critical point carries 26 scalar fields (ρ, $\nabla^2\rho$, ellipticity, kinetic/potential energy densities); ring and cage CPs, non-nuclear attractors, and atomic basin integrals are stored where present.

Fuzzy and surface descriptors. Fuzzy atom analysis partitions molecular properties via smooth, overlapping atomic weight functions, yielding basis-set-independent atomic populations, valences, and bond orders (Mayer & Salvador, 2004). Per-atom Becke and Hirshfeld fuzzy electron-density integrations (Becke, 1988) are computed via Multiwfn for every structure (with spin-density counterparts for open-shell records). These provide atomic electron and spin populations that are independent of the one-electron basis. Molecular-level ALIE (Average Local Ionization Energy) surface properties, including surface volume, area and extrema, and molecular planarity descriptors are additionally extracted per molecule, encoding site-specific electrophilicity and reactivity complementary to the atomic-level descriptors.

ORCA globals. Total SCF energy, HOMO/LUMO orbital energies, molecular dipole, nuclear gradient, and SCF convergence metadata are parsed from the ORCA output file (Neese et al., 2022; Neese, 2012).

Targets and diagnostics. Table 2 states the intended role of each descriptor family and the question it opens. "Auxiliary" describes intended downstream use, not exclusion: T1 regresses all six charge schemes jointly and weights them equally in its headline metric. Mulliken and Löwdin charges are deterministic functions of the density matrix in the fixed def2-TZVPD basis, so they are well-defined labels even where they disagree with the real-space schemes (App. J).

Table 2: Intended role of each descriptor family.

Family	Role	Question it opens
Hirshfeld, CM5, ADCH, Becke charges	Primary target (T1)	Can a surrogate replace real-space density partitioning at scale?
Mulliken, Löwdin charges	Auxiliary target (T1)	How far do basis-dependent populations track the real-space schemes?
QTAIM BCP scalars (ρ, $\nabla^2\rho$, ϵ)	Primary target (T2)	Can bond-topology descriptors be predicted without a wavefunction?
QTAIM BCP set	Primary target (T3)	Which atom pairs are bonded in the electron density?
Mayer, Löwdin bond orders	Reference (T3)	Do density-matrix bond orders agree with density topology on what a bond is?
ALIE surface descriptors	Secondary target	Can reactivity-relevant surface properties be predicted at scale?
ORCA globals	Provenance / filter	Quality control and conditioning, not targets.
`energy.lmdb`	Reference labels	Joint training of descriptor heads with OMol25 energies and forces.

3.2 Storage Layout

Each split folder contains nine LMDB files keyed on a single string identifier k = `relpath(folder, root).replace(os.sep, "__")`. Keys for each structure are shared across all lmdbs and they each contain different datatypes:

- `structure.lmdb`: pymatgen (Ong et al., 2013) `Molecule` (positions, elements), total spin and charge, RDKit-derived (Landrum et al., 2026) bond list, and external identifiers.

- `charge.lmdb`: per-atom charges nested by scheme (`hirshfeld`, `cm5`, `adch`, `becke` (ADC-corrected), `mulliken_orca`, `loewdin_orca`); `hirshfeld`/`cm5` also carry a total dipole, `adch`/`becke` carry total and per-atom dipoles.
- `qtaim.lmdb`: per-structure dict of critical points keyed by atom ID for nuclear CPs or by atom-pair for bond CPs; 26 scalar fields per CP including ρ, $\nabla^2\rho$, ellipticity, kinetic and potential energy densities.
- `bond.lmdb`: bond-order schemes (`fuzzy_bond`, `mayer_orca`, `loewdin_orca`) keyed by undirected atom-pair strings `"{i}_{El_i}_to_{j}_{El_j}"`; the fourth scheme, QTAIM bond presence, is recoverable from `qtaim.lmdb` BCP keys; index and threshold conventions are in App. E.
- `fuzzy.lmdb`: per-atom Multiwfn fuzzy integrations (`becke_fuzzy_density`, `hirsh_fuzzy_density`; spin counterparts for open-shell records).
- `other.lmdb`: ALIE surface statistics (surface volume and area, ALIE extrema, skewness) and four molecular planarity descriptors (`mpp_full`, `sdp_full`, `mpp_heavy`, `sdp_heavy`: planarity parameter and span of deviation from plane, all atoms and heavy atoms).
- `orca.lmdb`: ORCA-derived globals (total energy, orbital energies, dipole, gradient, SCF metadata), filtered via `DEFAULT_ORCA_FILTER` when consumed by the converter.
- `timings.lmdb`: per-job step durations; provenance only.
- `energy.lmdb`: OMol25 reference labels aligned to the same key space: total energy (eV), per-atom forces (eV/Å), the source OMol25 record identifier, and OMol25's per-record metadata.

Energies and forces. `energy.lmdb` joins the OMol25 total energies, forces and per-record metadata to the same keys for all 3,986,738 catalogued structures, including the jobs absent from the descriptor LMDBs. Against the SCF energies and gradients we parse independently into `orca.lmdb`, every shared record agrees to within 2.9×10^{-11} eV and 7.1×10^{-15} eV/Å (App. I).

3.3 Data Quality and Per-Vertical Breakdown

Calculations and conversions to LMDBs are highly stable with five of the seven analytic LMDBs showing no corrupt or incomplete data storage. For `qtaim` and `orca` we have 25 (0.0006%) and 1,607 (0.0403%) failed calculations; the ORCA failures concentrate on physically hard cases (solvated biomolecules, scaled-separation electrolytes, open-shell TM redox). The failed QTAIM records were instances where no BCPs were detected or where the Poincare-Hopf formula was not obeyed. The two failure sets overlap in one job, so 1,631 jobs failed, and each is removed from all eight descriptor LMDBs rather than left as a partial record. The descriptor LMDBs therefore hold 3,985,107 of the 3,986,738 catalogued structures, and `energy.lmdb` holds all of them. Because every LMDB uses the same key, a consumer joins families by key intersection: the eight descriptor LMDBs agree to within 7 keys, and joining all nine yields 3,985,099 structures (App. H, Table 8). Each of the six charge schemes (Hirshfeld, ADCH, CM5, Becke, Mulliken, Löwdin), four bond-order schemes (fuzzy, Mayer, Löwdin, QTAIM), full QTAIM topology, and the standard ORCA globals reaches $\geq$99% on every one of the 34 verticals; in other words, the same profile is available for a noble-gas dimer, a 200-atom solvated protein pocket, and a transition-metal redox complex. Full audit table, coverage heatmap, and per-vertical record counts: App. H and Table 5.

Census. Before the train/val/test partition, the corpus indexes $\sim$1.3M unique compositions and carries $\sim$218M atom-level records, over 30M charge-scheme records and $\sim$1.83 billion bond-scheme records across the four bond-order schemes. The schemes report different atom-pair sets, so bond counts are records, not unique bonds. Hirshfeld–CM5 is the most concordant charge pair corpus-wide ($R^2 = 0.77$), while Mulliken and Löwdin diverge most from the stockholder-partitioned schemes. Per-vertical counts, the per-scheme bond breakdown (Fig. 2) and the pairwise charge R^2 and MAE matrices are in App. D and App. J.

Cross-package reproducibility. Recomputing charges with HORTON (Chan et al., 2024) and QTAIM topology with Critic2 (Otero-de-la Roza et al., 2014) from the same wavefunctions for 87 structures across seven verticals, Hirshfeld charges agree to a median of 0.004 e ($r = 0.9993$),

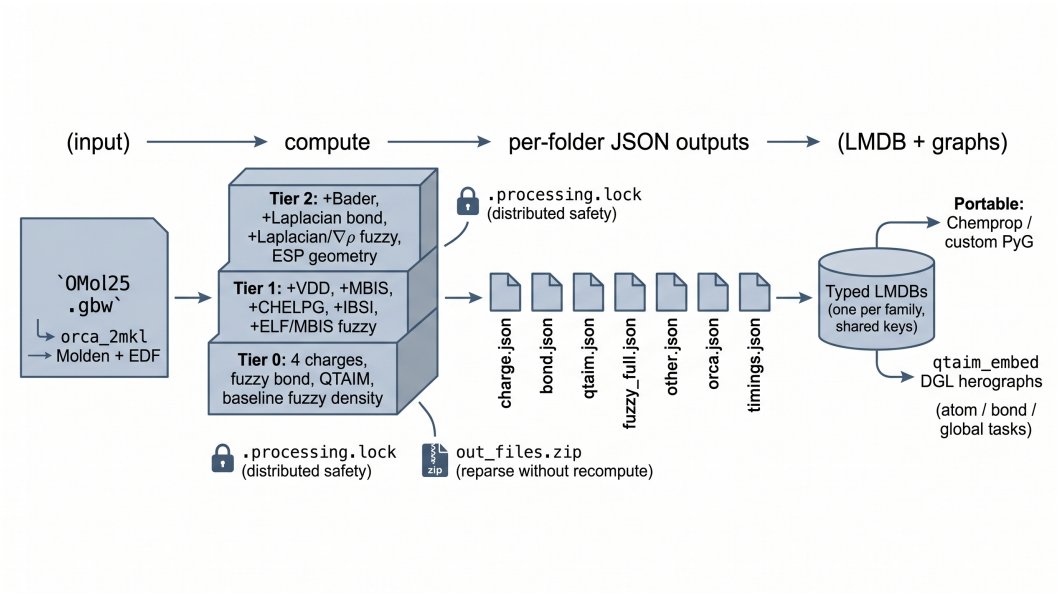

Figure 1: Overview of the `qtaim_generator` post-processing pipeline. Starting from an OMol25 `.gbw` wavefunction, the pipeline converts to Molden format (with EDF loading for ECP correctness), runs one of three calculation tiers through Multiwfn, and emits seven per-folder JSON outputs. These are converted to typed LMDBs sharing a common key, which can be consumed directly by graph-based frameworks or converted to PyG heterographs via our sister package `qtaim_embed` (`https://anonymous.4open.science/r/qtaim_embed-0E05`).

BCP properties agree with $r \geq 0.998$, and the BCP sets match exactly in 77 of 87 structures, with differences confined to ECP structures (App. M). This tests the software axis at a fixed wavefunction; sensitivity to level of theory is studied for QTAIM features by Gee et al. (2025).

4 PIPELINE

The pipeline is an orchestration layer over two external binaries, ORCA's `orca_2mkl` (Neese, 2012) and Multiwfn (Lu & Chen, 2012b), applied to the 4M OMol25 electronic-structure release that ships wavefunctions, density matrices, and raw ORCA outputs (`.out`, `.engrad`, `orca_property.txt`). For each job folder we read the `.gbw` and `.out`. Multiwfn does not read ORCA's binary `.gbw`, so the wavefunction is converted to Molden with `orca_2mkl`. For atoms treated with effective core potentials (ECPs) a Molden file describes only the valence density; Multiwfn restores the missing core density from its built-in EDF (electron density function) library (Lu & Chen, 2012b), so charges and QTAIM topology around heavy atoms see an all-electron-like density. After each run the pipeline checks the archived Multiwfn output for a completed EDF load and re-queues folders where it failed. A single folder runs with one command (App. P).

A single integer `full_set` $\in \{0, 1, 2\}$ selects a calculation tier. Tiers are strict supersets, and the QTAIM critical-point search and ORCA-derived globals run at every tier (Table 14, App. O); this release uses Tier 0. Each job folder carries a lock file and an archive of every Multiwfn output stream, so distributed workers do not collide and a folder can be reparsed without recomputing. `json-to-lmdb` converts the per-folder JSON outputs into typed LMDBs, one per descriptor family, sharing a key derived from the job folder's relative path. These are usable directly by external graph models such as Chemprop (Heid et al., 2024), or composed into PyG heterographs by `qtaim_embed`. The same code runs under PBS Pro, Slurm and Flux through Parsl (Babuji et al., 2019) and on a single workstation (App. O).

5 EVALUATION PROTOCOL

5.1 TASKS

This version reports baseline results for T1. T2 and T3 are specified below as complete protocols (targets, splits, metrics, stratification and reference baselines); their results will follow in a later version.

- **T1 – Per-atom partial charge regression.** Multi-target: ADCH, Becke, CM5, Hirshfeld, Löwdin and Mulliken (six schemes) jointly per atom. Headline: per-method MAE on the test split and held-out sets. Net charge and spin multiplicity enter as global input features; predicted charges are not constrained to sum to the net charge.
- **T2 – BCP property regression.** Targets: ρ, $\nabla^2\rho$, ellipticity at every QTAIM bond critical point (`density_all`, `lap_e_density` and `ellip_e_dens` in `qtaim.lmdb`). The setting is oracle: the BCP set, and therefore the bond graph, is taken from the reference QTAIM topology, and only the BCP properties are predicted. Bond paths are nodes of the heterograph and the loss is applied to bond nodes only. Headline: per-element-pair MAE and Pearson r for each target, stratified by bond class (below). SchNet4AIM (Gallegos et al., 2024), which regresses real-space QTAIM descriptors, is the comparator architecture.
- **T3 – Bond classification.** Binary: does a QTAIM BCP exist between an atom pair within $2\times$ sum-of-covalent-radii. Candidate pairs are all atom pairs within that cutoff, using RDKit covalent radii (Landrum et al., 2026), and labels are the BCP keys in `qtaim.lmdb`. The $2\times$ cutoff keeps the task non-trivial: on the reference sample below, 42.7 % of candidates are positive, against 82.6 % at $1.4\times$. Comparators are Mayer and Löwdin bond orders from `bond.lmdb`, binarized at ≥ 0.5. Headline: macro-F1 plus per-element-pair confusion statistics, always reported next to a geometric reference: a one-parameter distance rule (bonded if $d \leq k \sum r_{\mathrm{cov}}$, with $k = 1.1$ tuned) reaches macro-F1 0.971 on 105,618 candidate pairs from 800 `tm_react` structures. A learned T3 model has to beat this rule to show value beyond interatomic distance.

T2 bond classes. Each BCP is assigned one class from its element pair and density. *Metal-ligand*: either atom is a transition metal, lanthanide or actinide. *H-bond*: exactly one atom is H, the other is N, O, F, S or Cl, and $0.002 \leq \rho \leq 0.034$ a.u., the density range of the Koch–Popelier hydrogen-bond criteria (Koch & Popelier, 1995). *Covalent*: all remaining BCPs; this class also absorbs weak contacts that involve no hydrogen. The rule needs only element labels and the stored ρ, so it is defined for every structure in the corpus. A single / double / aromatic taxonomy would require a sanitized covalent bond graph, which is undefined for much of the corpus (App. G).

5.2 MAIN SPLIT

The corpus is partitioned by molecular composition: each structure's composition string – the electronegativity-ordered formula produced by `pymatgen`'s (Ong et al., 2013) `Composition.formula` – is hashed with SHA-256, mapped to $[0, 1)$, and assigned to train / val / test by cumulative thresholds $0.8/0.1/0.1$. Because the key is an exact function of composition, every structure sharing a composition receives the same assignment. Held-out sets (§5.3) are excluded prior to hashing, so no structure appears as both an in-distribution test point and a stress-test point. Hashing on composition rather than on row identity prevents near-duplicate conformers of the same formula from crossing the train/val/test boundary. Realized sizes are 3,125,019 / 398,982 / 403,951 (train / val / test) after excluding 58,786 held-out structures, which together account for the full 3,986,738-structure corpus. Realized ratios are 79.6 % / 10.2 % / 10.3 % of the post-holdout corpus, within 0.4 percentage points of the nominal targets. Verticals are not sampled uniformly because the split is global by formula; a per-vertical post-split distribution table is provided in App. F as a transparency artifact.

No test structure shares its exact composition with any training structure (0.0 %, against 80.9 % for a row-level random split of the same corpus). Scaffold- and homologue-level similarity is quantified in App. G.

5.3 HELD-OUT EVALUATION SETS

Held-out sets are constructed independently of the main split. Membership is decided from the manifest (and from `bond.lmdb` for the two metal-ligand sets) and removed from train/val/test so no structure is evaluated as both in-distribution test and stress-test. Subsampling is deterministic: H3 quotas are filled by a composition-stratified BLAKE2b (Aumasson et al., 2013) rank over `formula_hill`, and the H1 and H6 pairs are drawn from log-spaced frequency bands with fixed NumPy seeds (App. C). Overlaps between held-out sets are allowed (a structure can be both H7-large and H8-high-charge), but each set is generated separately so error attributes to one axis. The headline signal per set is the gap $\Delta = \mathrm{MAE}_{\text{held-out}} - \mathrm{MAE}_{\text{main test}}$, reported alongside the absolute errors. Δ is defined on MAE only; Pearson r, R^2 and F1 are reported as absolute values, since for them a drop, not a rise, signals degradation. The reference is the whole main test split for every suite: the pairs held out in H1 and H6 are purged from train, val and test, so no main-test subset contains them. A small Δ on H7 suggests accuracy transfers to larger systems; a large Δ on H1 or H6 indicates that accuracy depends on having seen the specific metal-partner pair. Three of the five sets (H1, H6, H7) inherit the hardest axes from the OMol25 paper directly.

Sets and sizes. The five held-out sets together contain 60,578 records spanning 58,786 distinct structures ($\sim$1.5% of the dataset; overlaps allowed, and 1,792 records are structures belonging to two sets):

- **H1 – Metal-ligand pairs (15,025):** eleven (TM, partner) bond pairs sampled stratified across log-spaced frequency bands; selection driven by *true* bonds (Mayer/fuzzy ≥ 0.3) from `bond.lmdb`, not element co-occurrence (App. C.1).
- **H3 – Reactivity (12,506):** composition-stratified subsample of `tm_react` (5k), `electrolytes_reactivity` (5k), `pmechdb` (2.5k); tests non-equilibrium and transition-state-adjacent geometries.
- **H6 – Lanthanide-ligand pairs (2,589):** same construction as H1 applied to (Ln, partner) bonds (App. C.2).
- **H7 – Large systems (18,096):** `n_atoms > 250`; size extrapolation where fuzzy integration and grid-based QTAIM are most expensive and per-atom DFT SNR is hardest to maintain.
- **H8 – Large net charges (12,362):** `net_charge_abs > 4`; the high-charge tail where partial-charge schemes disagree most and SCF is most sensitive to the basis set.

Counts are the records present in the released held-out LMDBs, verified against each database's stored length. They fall a few records short of the selection targets (H1 15,030, H3 12,507, H7 18,200, H8 12,393) where a selected structure lacked a complete descriptor set at pull time. H6 has no fixed selection count; its sampling shape targets $\sim$2,500 (App. C.2).

Because membership is decided per structure and the main split per composition, some held-out structures share their composition with a training structure, from 0.9 % (H7) to 21.7 % (H1); H3 frames that were not sampled can hash into train. We therefore report each suite on its full membership and on its formula-disjoint subset (App. C.3).

We deliberately use pair-level held-outs (H1, H6) rather than whole-element held-outs because the goal of this benchmark is to produce models that are usable across the periodic table. Holding out specific (metal, ligand) pairs keeps every element represented in training while still measuring rare-chemistry transfer, identical to OMol25.

5.4 METHODOLOGY AND METRICS

All hyperparameter tuning and early stopping use the validation bucket; the test bucket is reserved for a single final evaluation, with the same one-shot rule on each held-out set. Metrics are MAE, RMSE, and Pearson r per task, plus R^2 for T1. Stratifications: **T1** per-method; **T2** per-element-pair, per-bond-class (covalent / metal-ligand / H-bond), per-vertical; **T3** macro-F1 plus per-element-pair confusion.

5.5 T1 Baseline Results

We train five models on T1, all predicting the six schemes jointly: a graph convolutional network (GCN) and a graph attention network (GAT) on the structural graph alone, and the same GCN trunk with a SchNet (Schütt et al., 2018), DimeNet++ (Gasteiger et al., 2020) or NequIP-style E(3)-equivariant tensor-product (Batzner et al., 2022; Geiger & Smidt, 2022) geometric encoder. The models were trained on an earlier assignment of the composition split that differs from the released split in 48,791 records (1.2 %); they are evaluated on its 403,484 test structures and on the five held-out sets, which are identical in both assignments. Counts in Table 3 are structures with a graph: 35 H7 and 39 H8 records have none (non-finite charges or a failed graph build). Table 3 reports MAE averaged over schemes and the gap Δ to each model's own test MAE. The equivariant encoder has the lowest MAE on every set (0.039 e on test, 0.036–0.055 e held-out), and the three geometric encoders reduce test MAE from 0.062 e (GCN) to 0.039–0.047 e. Per-scheme MAE and R^2 are in App. L.

Table 3: T1 mean MAE over the six charge schemes (e), with $\Delta = \mathrm{MAE}_{\mathrm{set}} - \mathrm{MAE}_{\mathrm{test}}$ per model in parentheses, and mean R^2 over schemes on test.

Set	N	GCN	GAT	SchNet	DimeNet++	Equivariant
Test	403,484	0.062	0.074	0.047	0.044	0.039
H1	15,025	0.082 (+0.020)	0.103 (+0.028)	0.067 (+0.021)	0.063 (+0.019)	0.055 (+0.016)
H3	12,506	0.063 (+0.001)	0.084 (+0.010)	0.055 (+0.009)	0.057 (+0.013)	0.046 (+0.007)
H6	2,589	0.071 (+0.009)	0.091 (+0.017)	0.060 (+0.014)	0.056 (+0.012)	0.049 (+0.010)
H7	18,061	0.076 (+0.015)	0.081 (+0.006)	0.041 (−0.005)	0.038 (−0.006)	0.036 (−0.003)
H8	12,323	0.065 (+0.003)	0.076 (+0.002)	0.044 (−0.003)	0.042 (−0.003)	0.038 (−0.001)
Test R^2		0.813	0.775	0.854	0.892	0.904

6 Limitations

This dataset is a subset of OMol25; release of the full electronic-structure slice would scale us into the hundreds of billions of atom and bond records, but we already cover the original's rich set of verticals. Hirshfeld, Becke and QTAIM descriptors are cross-checked against HORTON and Critic2 on an 87-structure subset (App. M), but ADCH, ADC-Becke, CM5 and the fuzzy bond orders have a single implementation (Multiwfn), and we did not benchmark against the commercial AIMAll (Keith, 2019). Compute budget restricted us to the Tier 0 features; future releases will add the more expensive schemes relevant to other communities (MBIS for force-field (Verstraelen et al., 2016) and MLIP development (Gong et al., 2025), IBSI for bond-strength analysis (Klein et al., 2020)). The original work's energy and force labels are merged into the release as `energy.lmdb`, aligned to the same per-structure keys across all splits and held-out suites, enabling partial-charge-aware MLIP development.

We are bound by the single level-of-theory used in OMol25 and so do not measure descriptor sensitivity to basis set or functional. Prior work by Gee et al. (2025) found that QTAIM features can vary significantly across LOTs while remaining useful for ML. OMol25 also covers only ground-state calculations, so excited-state sensitivity is out of scope.

7 Conclusion

We present OMol-Descriptors-4M: six partial charges, four bonding schemes, full QTAIM topology, fuzzy descriptors, and ORCA globals across ∼4M structures and 34 verticals, paired with a portable post-DFT pipeline and a held-out evaluation protocol. This lowers the barrier for descriptor-aware surrogate models across catalysis, drug discovery, and electrolyte design.

AI use statement

In this work, we used generative AI tools (Claude Code) to implement methods: parts of the `qtaim_generator` pipeline code, the LMDB conversion, audit and cleanup scripts, and the

analysis and figure scripts were written or edited with AI assistance. We also used generative AI tools to help design and run dataset-cleaning steps (removing failed jobs and a duplicated charge scheme from the LMDBs, realigning split assignments). We have not used generative AI tools to generate synthetic data, to formulate hypotheses or theoretical models, or to translate text; mathematical claims and proofs are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the manuscript for readability, to format references, and to obtain an automated pre-submission review of an earlier draft. We have reviewed all AI-assisted work: AI-written code was checked against the repository test suite and by the authors, and every number reported in the paper was traced to the output of a script run by the authors on the released LMDBs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

OMol-Descriptors-4M is derived entirely from the public OMol25 electronic-structure release and inherits its CC-BY-4.0 license terms. The work involves no human subjects, personal data, or crowdsourcing. The dual-use risk of this release is minimal. The dataset surfaces interpretive electronic-structure descriptors (charges, bond orders, QTAIM topology) on neutral and charged molecules computed from wavefunctions that are already public, and it confers no obvious uplift for chemical, biological, or materials threats beyond what the underlying OMol25 wavefunctions already permit. We release the data and pipeline under an open license to support reproducibility in chemistry ML. We note the known data-quality limitations of the release in Section 6 and the datasheet (App. A) so that downstream users do not over-read the labels.

Reproducibility statement

The anonymized pipeline code is available at `https://anonymous.4open.science/r/qtaim_generator-8C0C/README.md` and the graph-construction package at `https://anonymous.4open.science/r/qtaim_embed-0E05`. The validation and test splits and all five held-out suites are available at `https://huggingface.co/datasets/confsubanon/OMol-Descriptors-4M`. The pipeline, its calculation tiers and HPC configurations are described in Section 4 and App. O, with a single-folder example in App. P. The storage layout and key scheme are in Section 3.2 and App. E; the split and held-out construction are in Section 5 and App. C; the conversion audit and cross-package validation are in App. H and App. M; the compute resources used are in App. B.

References

Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018. doi: 10.1126/science.aar5169.

Dylan M. Anstine, Roman Zubatyuk, and Olexandr Isayev. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science*, 16(23): 10228–10244, 2025. doi: 10.1039/D4SC08572H.

Jean-Philippe Aumasson, Samuel Neves, Zooko Wilcox-O'Hearn, and Christian Winnerlein. Blake2: Simpler, smaller, fast as md5. In Michael Jacobson, Michael Locasto, Payman Mohassel, and Reihaneh Safavi-Naini (eds.), *Applied Cryptography and Network Security*, pp. 119–135, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-38980-1.

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9:185, 2022.

Yadu Babuji, Anna Woodard, Zhuozhao Li, Daniel S. Katz, Ben Clifford, Rohan Kumar, Lukasz Lacinski, Ryan Chard, Justin M. Wozniak, Ian Foster, Michael Wilde, and Kyle Chard. Parsl: Pervasive parallel programming in Python. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing (HPDC '19)*, pp. 25–36. ACM, 2019. doi: 10.1145/3307681.3325400.

Richard F W Bader. *Atoms in Molecules: A Quantum Theory*. Oxford University Press, 1990.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. Open materials 2024 (OMat24) inorganic materials dataset and models, 2024. URL https://arxiv.org/abs/2410.12771.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, et al. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18): 184110, 2025. doi: 10.1063/5.0297006.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5.

A D Becke. A multicenter numerical integration scheme for polyatomic molecules. *The Journal of Chemical Physics*, 88(4):2547–2553, 2 1988. doi: 10.1063/1.454033.

Eric Berquist, Amanda Dumi, Shiv Upadhyay, Omri D. Abarbanel, Minsik Cho, Sagar Gaur, Renee Gil, Geoffrey R. Hutchison, Oliver S. Lee, Andrew S. Rosen, Sanjeed Schamnad, Felipe S. S. Schneider, Casper Steinmann, Maxim Stolyarchuk, Jonathon E. Vandezande, Weronika Zak, and Karol M. Langner. cclib 2.0: An updated architecture for interoperable computational chemistry. *The Journal of Chemical Physics*, 161(4):042501, 07 2024. ISSN 0021-9606. doi: 10.1063/5.0216778. URL https://doi.org/10.1063/5.0216778.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners, 2020.

Matthew Chan, Toon Verstraelen, Alireza Tehrani, Michelle Richer, Xiaotian Derrick Yang, Taewon David Kim, Esteban Vöhringer-Martinez, Farnaz Heidar-Zadeh, and Paul W. Ayers. The tale of HORTON: Lessons learned in a decade of scientific software development. *The Journal of Chemical Physics*, 160(16):162501, 2024. doi: 10.1063/5.0196638.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, et al. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

Christian Devereux et al. Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *Journal of Chemical Theory and Computation*, 16:4192–4202, 2020.

Peter Eastman et al. SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10:11, 2023.

Miguel Gallegos, José Jiménez-Luna, et al. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-48567-9.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *Machine Learning for Molecules Workshop, NeurIPS*, 2020.

Winston Gee, Abigail Doyle, Santiago Vargas, and Anastassia N. Alexandrova. Multi-level qtaim-enriched graph neural networks for resolving properties of transition metal complexes. *Digital Discovery*, 4:3378–3388, 2025. doi: 10.1039/D5DD00220F.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Sheng Gong, Yumin Zhang, Zhenliang Mu, Zhichen Pu, Hongyi Wang, Xu Han, Zhiao Yu, Mengyi Chen, Tianze Zheng, Zhi Wang, Lifei Chen, Zhenze Yang, Xiaojie Wu, Shaochen Shi, Weihao Gao, Wen Yan, and Liang Xiang. A predictive machine learning force-field framework for liquid electrolyte development. *Nature Machine Intelligence*, 7(4):543–552, 2025. doi: 10.1038/s42256-025-01009-7.

Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: A machine learning package for chemical property prediction. *J. Chem. Inf. Model.*, 64(1):9–17, January 2024.

F L Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theoretica Chimica Acta*, 44(2):129–138, 1977. doi: 10.1007/BF00549096.

Johannes Hoja, Leonardo Medrano Sandonas, Brian G Ernst, et al. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific Data*, 8:43, 2021. doi: 10.1038/s41597-021-00812-2.

Matthew K Horton, Patrick Huck, Ruo Xi Yang, et al. Accelerated data-driven materials science with the Materials Project. *Nature Materials*, 24(10):1522–1532, 2025. doi: 10.1038/s41563-025-02272-0.

Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. QMugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9:273, 2022. doi: 10.1038/s41597-022-01390-7.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Todd A Keith. AIMAll (version 19.10.12). TK Gristmill Software, Overland Park KS, USA, 2019. `aim.tkgristmill.com`.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, et al. Segment anything, 2023.

Johanna Klein, Hassan Khartabil, Jean-Charles Boisson, Julia Contreras-García, Jean-Philip Piquemal, and Eric Hénon. New way for probing bond strength. *The Journal of Physical Chemistry A*, 124 (9):1850–1860, 2020. doi: 10.1021/acs.jpca.9b09845.

Ulrich Koch and Paul L. A. Popelier. Characterization of C-H-O hydrogen bonds on the basis of the charge density. *The Journal of Physical Chemistry*, 99(24):9747–9754, 1995. doi: 10.1021/j100024a016.

Greg Landrum et al. RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2026.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The open molecules 2025 (omol25) dataset, evaluations, and models, 2026. URL `https://arxiv.org/abs/2505.08762`.

Shih-Cheng Li, Haoyang Wu, Angiras Menon, Kevin A. Spiekermann, Yi-Pei Li, and William H. Green. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? *Journal of the American Chemical Society*, 146(33):23103–23120, 2024. doi: 10.1021/jacs.4c04670.

Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. Highly accurate quantum chemical property prediction with Uni-Mol+, 2023.

Tian Lu and Feiwu Chen. Atomic dipole moment corrected Hirshfeld population method. *Journal of Theoretical and Computational Chemistry*, 11(01):163–183, 2012a. doi: 10.1142/S0219633612500113.

Tian Lu and Feiwu Chen. Multiwfn: A multifunctional wavefunction analyzer. *Journal of Computational Chemistry*, 33:580–592, 2012b.

Aleksandr V Marenich, Steven V Jerome, Christopher J Cramer, and Donald G Truhlar. Charge model 5: An extension of Hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases. *Journal of Chemical Theory and Computation*, 8(2): 527–541, 2012. doi: 10.1021/ct200866d.

I Mayer. Charge, bond order and valence in the AB initio SCF theory. *Chemical Physics Letters*, 97 (3):270–274, 1983. doi: 10.1016/0009-2614(83)80005-0.

I Mayer and P Salvador. Overlap populations, bond orders and valences for 'fuzzy' atoms. *Chemical Physics Letters*, 383(3):368–375, 2004. doi: 10.1016/j.cplett.2003.11.048.

Brandon Meza-González, David I. Ramírez-Palma, Pablo Carpio-Martínez, David Vázquez-Cuevas, Karina Martínez-Mayorga, and Fernando Cortés-Guzmán. Quantum topological atomic properties of 44K molecules. *Scientific Data*, 11(1):945, 2024. doi: 10.1038/s41597-024-03723-0.

H L Morgan. The generation of a unique machine description for chemical structures – a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi: 10.1021/c160017a018.

F. Neese. The orca program system. *WIRES Comput. Molec. Sci.*, 2(1):73–78, 2012. doi: 10.1002/wcms.81.

Frank Neese et al. Software update: The ORCA program system – version 5.0. *WIREs Computational Molecular Science*, 12:e1606, 2022.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.

A Otero-de-la Roza, Erin R Johnson, and Víctor Luaña. Critic2: A program for real-space analysis of quantum chemical interactions in solids. *Computer Physics Communications*, 185(3):1007–1018, 2014. doi: 10.1016/j.cpc.2013.10.026.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, et al. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 2020. doi: 10.3389/fphar.2020.565644.

Jay W. Ponder, Chuanjie Wu, Pengyu Ren, Vijay S. Pande, John D. Chodera, Michael J. Schnieders, Imran Haque, David L. Mobley, Daniel S. Lambrecht, Robert A. DiStasio, Martin Head-Gordon, Gary N. I. Clark, Margaret E. Johnson, and Teresa Head-Gordon. Current status of the AMOEBA polarizable force field. *The Journal of Physical Chemistry B*, 114(8):2549–2564, 2010. doi: 10.1021/jp910674d.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Mathias Schreiner, Arghya Bhowmik, et al. Transition1x – a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9:779, 2022.

K T Schütt, H E Sauceda, P-J Kindermans, A Tkatchenko, and K-R Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

Justin S Smith et al. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data*, 7:134, 2020.

Peter C St. John, Yanfei Guan, Yeonjoon Kim, Seonah Kim, and Robert S Paton. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nature Communications*, 11:2328, 2020. doi: 10.1038/s41467-020-16201-z.

Annika Stuke et al. A dataset of 62,000 organic molecules with first-principles properties. *Scientific Data*, 7:58, 2020.

Alireza Tehrani, Xiaotian Derrick Yang, Marco Martínez-González, Leila Pujal, Raymundo Hernández-Esparza, Matthew Chan, Esteban Vöhringer-Martinez, Toon Verstraelen, Paul W. Ayers, and Farnaz Heidar-Zadeh. GRID: A Python library for molecular integration, interpolation, differentiation, and more. *The Journal of Chemical Physics*, 160(17):172503, 2024. doi: 10.1063/5.0202240.

Santiago Vargas, Matthew R. Hennefarth, Zhihao Liu, and Anastassia N. Alexandrova. Machine learning to predict Diels–Alder reaction barriers from the reactant state electron density. *Journal of Chemical Theory and Computation*, 17(10):6203–6213, 2021. doi: 10.1021/acs.jctc.1c00623.

Santiago Vargas, Winston Gee, and Anastassia N. Alexandrova. High-throughput quantum theory of atoms in molecules (QTAIM) for geometric deep learning of molecular and reaction properties. *Digital Discovery*, 3(5):987–998, 2024. doi: 10.1039/D4DD00057A.

Toon Verstraelen, Steven Vandenbrande, Farnaz Heidar-Zadeh, Louis Vanduyfhuys, Veronique Van Speybroeck, Michel Waroquier, and Paul W Ayers. Minimal basis iterative stockholder: Atoms in molecules for force-field development. *Journal of Chemical Theory and Computation*, 12(8):3894–3912, 2016. doi: 10.1021/acs.jctc.6b00456.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025. doi: 10.1038/s41586-025-08628-5. URL https://doi.org/10.1038/s41586-025-08628-5.

A Datasheet for Datasets

Motivation. OMol-Descriptors-4M was created to extend the OMol25 corpus (Levine et al., 2026) with post-DFT interpretive descriptors: partial charges, bond orders, QTAIM topology, fuzzy and surface integrations, and ORCA global features (which were in OMol25) that have historically been absent from large-scale ML-ready quantum-chemical datasets. The dataset is intended for training descriptor-aware ML models and for benchmarking surrogates that approximate wavefunction-derived quantities without per-molecule DFT. The dataset was created by the authors at the authors' home institution. Funding sources are listed in the Acknowledgments.

Composition. Each instance is a single molecular structure with associated wavefunction-derived descriptors. The catalogue contains 3,986,738 structures, the full public 4M wavefunction subset of OMol25, spanning 80+ elements, and `energy.lmdb` carries energies and forces for all of them. Descriptor LMDBs cover 3,985,107: the 1,631 jobs whose QTAIM or ORCA post-processing failed (Table 7) are absent from every descriptor LMDB rather than partially populated. The eight descriptor LMDBs then agree to within 7 keys (Table 8), and an inner join of all nine LMDBs on the shared key yields 3,985,099 structures. Per-structure data is sharded across nine LMDB files (eight descriptor LMDBs plus `energy.lmdb`) (Section 3.2); the descriptor families are enumerated in Section 3. Splits are described in Section 5. The corpus contains no human subjects, personally identifiable information, or confidential data.

Collection Process. Geometries, wavefunctions, density matrices, and ORCA output files were inherited from the OMol25 4M electronic structure release. We did not collect new DFT calculations; all post-processing was applied to artifacts already produced by the OMol25 authors at the ωB97M-V/def2-TZVPD level.

Preprocessing / Cleaning / Labeling. Wavefunctions were processed through the open `qtaim_generator` pipeline (Section 4), wrapping Multiwfn (Lu & Chen, 2012b) and ORCA (Neese et al., 2022) post-analysis utilities. JSON outputs were validated and compacted into LMDB shards keyed by relative job-folder path. Conversion is essentially lossless (Section 3.3); all raw artifacts remain available in the OMol25 release for users wishing to reprocess.

Uses. Intended uses include training ML models that predict charges, bond orders, or QTAIM properties; benchmarking descriptor-aware MLIPs; and serving as a target for surrogate models that compress the full DFT-then-Multiwfn pipeline. Inappropriate uses include drawing chemistry conclusions outside the OMol25 level of theory or extrapolating to excited-state regimes (see §6).

Distribution. The dataset is released on Hugging Face under a CC-BY-4.0 license, inheriting OMol25's license terms. For review, an anonymous mirror hosts the validation and test splits and the five held-out suites with SHA-256 checksums: `https://huggingface.co/datasets/confsubanon/OMol-Descriptors-4M`. The train split is released with the de-anonymized version. The `qtaim_generator` pipeline is released under MIT license at `https://anonymous.4open.science/r/qtaim_generator-8C0C/README.md`. No third-party IP, export controls, or fees apply.

Maintenance. The dataset will be maintained by the authors. Issues, errata, and feature requests can be filed on the code repository. Updates, including reruns of the jobs that failed post-processing and additional descriptor schemes (MBIS, IBSI, CHELPG), will be released as versioned HF records with changelogs. Older versions will remain accessible.

B Compute Resources

Post-processing of the 4M-structure subset was carried out on three HPC systems; App. O lists where each scheduler configuration was exercised, which is a different set (names withheld for review):

- **System A** (Flux scheduler). Used for the highest-throughput phase of Multiwfn-driven QTAIM, charge, bond-order, and fuzzy-integration jobs.

- **System B** (PBS Pro, AMD EPYC nodes). Used for additional descriptor generation and Parsl-managed workflow batches.
- **System C** (Slurm). Used for ORCA conversions (`orca_2mkl`), JSON-to-LMDB sharded conversion, and analytical notebook workloads.

C Held-Out Construction Logic

C.1 H1 metal-ligand pair selection

- Element-pair co-occurrence (`metal_nonmetal_pairs.csv`, 859 TM–nonmetal pairs) covers nonmetal partners only and is not used for selection; selection here uses true bonds (Mayer / fuzzy bond order $\geq$ 0.3) extracted from `bond.lmdb` via `tm_neighbor_lists.py`.
- Per-vertical bond extraction across 15 TM-bearing verticals; one row per (TM atom, partner atom) bond.
- Global pair table: **1,101 unique (TM, partner) bond pairs**, including metal–metal pairs (e.g. Pt–Na, Rh–Fe, Ag–Pd), across **651,742 TM-bonded structures** (16.35% of dataset).
- Frequency bands (log-spaced): B1 $[1, 10)$ 529 pairs; B2 $[10, 100)$ 93; B3 $[100, 1k)$ 233; B4 $[1k, 10k)$ 203; B5 $[10k, 100k)$ 43.
- Stratified sampling shape $(3, 2, 2, 4, 0)$, `numpy.random.default_rng(seed=87)`. B5 excluded (Pt-C, Ti-H, etc. dominate training); B4 capped at 4 to keep realized holdout near 15k.
- Selected pairs: *B1* Pt-Na, Rh-Fe, Ag-Pd; *B2* Sc-Ir, Cd-Sb; *B3* Mn-Sb, Ir-Ge; *B4* Tc-Cl, Re-H, Sc-S, Hf-C.
- Variance: across 30 seeds with shape $(3, 2, 2, 4, 0)$, realized size is $15{,}227 \pm 4{,}871$. Seed 87 lands within 200 of the 15k target.
- Reproducibility: `qtaim_gen/source/analysis/build_holdout_csvs.py`; outputs `h1_metal_ligand_pair_definitions.csv` and `h1_metal_ligand_pairs.csv`.

C.2 H6 lanthanide-ligand pair selection

- Same construction as H1, applied to (Ln, partner) bonds via `tm_neighbor_lists.py --element_class ln`.
- Five Ln-bearing verticals: `omol`, `tm_react`, `ml_mo`, `electrolytes_reactivity`, `mo_hydrides`.
- Global pair table: **453 unique (Ln, partner) bond pairs** across **66,677 Ln-bonded structures** (1.67% of dataset). All 15 lanthanides represented.
- Frequency bands: B1=207, B2=102, B3=86, B4=58, B5=0 (max pair size 4,439). H6 effectively samples four bands.
- Stratified sampling shape $(3, 2, 2, 1, 0)$, `seed=190`. B4 toned down to 1 vs H1's 4 (Ln B4 is sparser, target $\sim$2.5k).
- Selected pairs: *B1* Lu-Ni, Ce-La, Gd-Sb; *B2* Ho-F, Yb-Si; *B3* Dy-S, Nd-S; *B4* Tm-N.
- Reproducibility: same build script as H1; outputs `h6_lanthanide_ligand_pair_definitions.csv` and `h6_lanthanide_ligand_pairs.csv`.

C.3 Composition overlap between held-out suites and train

Held-out membership is decided per structure, while the main split is decided per composition, so a held-out structure can share its composition with structures that hashed into train ("train sisters"). For H3 this is by design: suite quotas are sampled per structure, so frames of the same composition that were not sampled remain in the candidate pool. Table 4 measures the overlap

against the released train split. Within H3 it comes from `pmechdb` (14.1 % of its 2,502 structures) and `electrolytes_reactivity` (9.2 % of 5,004), not `tm_react` (0.1 % of 5,000), and the H3 train sisters come mostly from other verticals (`trans1x` 3,778, `ani1xbb` 3,085, `ani2x` 3,000, `rgd_uks` 1,727 rows): these are small compositions shared across verticals, not unsampled frames of the same reaction. The high H1 fraction reflects that H1 is defined by a metal-partner bond pair, which a structure can contain while sharing its composition with a training structure; H1 measures pair-level, not compositional, novelty. We report every suite on its full membership and on its formula-disjoint subset. Source: `scripts/holdout_sibling_leakage.py`.

Table 4: Held-out structures that share their composition with at least one structure in the released train split. *Disjoint*: structures with no train sister. A structure in two suites is counted in both.

Suite	Structures	Compositions	With train sister		Train sister rows	Disjoint
			Structures	Compositions		
H1	15,025	7,971	3,256 (21.7 %)	1,625 (20.4 %)	3,800	11,769
H3	12,506	5,537	815 (6.5 %)	242 (4.4 %)	14,262	11,691
H6	2,589	1,723	181 (7.0 %)	133 (7.7 %)	190	2,408
H7	18,096	13,443	158 (0.9 %)	31 (0.2 %)	31	17,938
H8	12,362	7,073	695 (5.6 %)	328 (4.6 %)	940	11,667

C.4 Holdout-Aware Splitting Pipeline

- **Step 1 – Pull holdouts.** `pull_holdout_records.py` reads each holdout CSV, materializes per-set LMDBs under `data/holdouts/<set>/`, and emits a `manifest_holdout.parquet` keyed on the LMDB key convention `relpath(folder, root_dir).replace(os.sep, "__")`.
- **Step 2 – Exclude.** `multi_vertical.load_exclusion_set` re-derives LMDB keys from the manifest `rel_path` column and unions H1/H3/H6/H7/H8 into one per-vertical exclusion set.
- **Step 3 – Split.** `split_descriptor_lmdbs.py` reads `structure.lmdb`, drops excluded keys, assigns each remaining key to train/val/test by `SHA-256(Composition.formula)`, mapped to $[0, 1)$ and thresholded at $0.8/0.1/0.1$; the composition string is `pymatgen`'s electronegativity-ordered formula, not Hill notation, then partitions each of the 8 descriptor LMDBs (`structure`, `charge`, `qtaim`, `bond`, `fuzzy`, `other`, `orca`, `timings`) into the assigned split.
- Output layout: `<output_dir>/<vertical>/<split>/<descriptor>.lmdb`.
- HPC parallelism: `--include_verticals X` + `--merge_from DIR`, mirroring the pull pipeline. Per-vertical `split_assignment.json` is the durable, race-free record (no global plan file).
- Lock-step audit: across all 8 descriptors, source row count = train + val + test + excluded for every vertical (verified on `ani2x`: 215,642 source = 168,783 + 17,417 + 29,442, $n_{\text{excluded}} = 0$).

D Per-Vertical Census Table

Table 5 expands Sec. 3.3 into the full per-vertical breakdown (34 verticals plus the corpus total). Figure 3 reports per-structure descriptor density for atoms and QTAIM bond critical points.

Table 5: Per-vertical census of OMol-Descriptors-4M. Counts are corpus-level pre-split totals. *Structs*: structure count. *Forms*: unique alphabetical formulas. *Atoms*: atom-level descriptor records. *Charge rec.*: per-job per-method charge-scheme records. *BCPs*: QTAIM bond critical points. *Mayer / Löwdin / Fuzzy*: per-scheme bond-order records. *Structs* counts structures in the descriptor LMDBs (3,985,107 of the 3,986,738 catalogued; App. H). The remaining `ml_mo` columns still include 23 failed jobs removed from the release, under 0.001 % of each column total.

Vertical	Structs	Forms	Atoms	Charge rec.	BCPs	Mayer	Löwdin	Fuzzy
5A_elytes	14,200	11,294	2,325,187	99,383	2,982,866	4,474,371	6,655,293	4,244,442
ani1xbb	737,208	48,320	10,834,251	7,318,728	10,801,957	12,465,497	30,634,609	28,118,840
ani2x	215,642	5,773	3,280,277	1,586,066	3,346,555	3,824,490	9,516,073	6,379,670
dna	24,647	4,312	2,186,892	172,529	2,559,225	3,843,725	7,372,380	4,907,003
droplet	18,694	3,376	1,109,505	130,858	1,457,592	1,653,393	2,805,247	1,785,875
electrolytes_reactivity	66,967	18,063	874,449	468,769	887,548	986,445	2,340,553	1,953,147
electrolytes_redox	46,069	33,579	4,017,583	322,483	4,945,260	7,442,261	11,512,752	9,064,929
electrolytes_scaled_sep	30,768	21,186	2,307,396	215,369	2,847,853	11,683,800	8,801,619	5,116,940
geom_orca6	199,805	36,054	10,321,888	1,398,634	11,579,867	18,310,972	35,343,998	22,839,614
low_spin_23	12,939	12,045	927,730	90,573	1,113,772	2,194,086	3,228,506	2,213,493
ml_elytes	51,560	35,054	7,608,642	360,904	9,672,793	15,638,665	22,494,066	13,464,794
ml_mo	77,580	40,658	6,098,677	543,153	7,336,479	12,931,862	20,795,278	17,937,539
ml_protein_interface	61,856	35,118	8,314,895	432,979	9,480,078	16,584,221	26,101,394	15,494,518
mo_hydrides	3,396	2,870	232,942	23,772	280,311	664,784	851,813	781,762
nakb	5,606	1,188	651,552	50,642	777,349	1,297,268	2,280,331	1,551,414
noble_gas	1,915	136	66,993	13,651	84,800	43,394	116,497	87,580
noble_gas_compounds	18	12	189	126	179	326	560	518
omol	1,428,279	593,876	92,071,794	9,997,947	109,334,787	170,492,089	291,413,475	218,967,571
orbnet_denali	51,102	30,339	2,305,764	357,714	2,594,095	4,028,552	7,875,696	5,188,774
pdb_fragments_300K	93,334	68,881	11,255,165	653,329	13,208,426	21,291,730	36,314,068	22,292,731
pdb_fragments_400K	92,698	68,485	11,177,181	648,881	13,058,825	20,774,760	36,149,798	22,215,370
pdb_pockets_300K	16,870	15,453	3,245,155	118,069	3,915,033	8,616,479	10,415,619	6,407,044
pdb_pockets_400K	3,204	3,092	630,613	22,422	701,104	1,658,125	2,019,313	1,222,919
pmechdb	49,832	28,171	2,951,782	379,418	3,316,002	6,028,189	9,928,066	6,572,474
protein_core	35,406	14,980	4,197,118	247,840	4,436,976	8,229,618	13,219,684	7,738,794
protein_interface	44,760	16,127	4,298,048	313,320	4,938,560	7,052,373	13,235,662	8,007,220
rgd_uks	75,316	980	1,337,124	527,212	1,340,336	1,504,737	3,810,675	3,314,638
rmechdb	2,803	731	58,528	28,024	59,403	76,928	171,675	150,328
rna	34,434	5,168	3,661,659	241,036	4,108,650	7,166,774	12,646,859	8,583,065
rpmd	18,652	2,992	1,412,713	152,365	1,721,533	2,025,287	3,893,696	2,478,985
scaled_separations_exp	53,180	37,168	5,020,630	372,260	5,549,452	5,214,145	13,616,510	8,822,261
spice	44,447	19,603	1,553,948	311,129	1,686,049	2,413,029	4,992,500	3,292,358
tm_react	158,189	86,819	9,209,119	1,107,323	10,873,610	17,179,722	29,890,848	22,389,065
trans1x	213,731	171	2,974,386	1,496,117	2,977,381	3,345,716	8,459,481	7,799,805
TOTAL	3,985,107	1,302,074	218,519,775	30,203,025	253,974,706	401,137,813	688,904,594	491,385,480

D.1 Sharded Execution and LMDB Merge

LMDB is single-writer and exhibits lock contention under naive parallel writes, so `json-to-lmdb` does not attempt to share a single output environment across workers. Instead, each parallel job is given a shard index $i \in \{0, \ldots, N-1\}$ and a total shard count N, and partitions the input folder list by index modulo N. For the flat *root/category/subset/job* hierarchy this is a glob plus a modulo filter; for jagged hierarchies such as OMol4M, where job folders sit at variable depth, the user supplies an explicit `--folder_list` file of absolute paths and the same modulo partitioning applies to the line index. In both cases the LMDB key is `relpath(folder, root_dir).replace(os.sep, "__")`, which is identical across every descriptor family, so the per-family shards always share keys and joining is trivial.

Each shard writes into its own subdirectory as `shard_{i}.lmdb`. Every shard writes a sentinel file (`shard_{i}.done`) when it finishes; the merge step polls until all N sentinels exist and then copies each shard's keys transactionally into a single consolidated LMDB per family. Folder order is `sorted()` glob order in glob mode and line order in `--folder_list` mode. The same shard/merge pattern is implemented at the converter stage in `converter.py`, so graph construction can also be parallelized and joined into one final per-vertical graph LMDB.

E Key and Index Conventions

Atom indices inside `charge.lmdb`, `bond.lmdb` and `qtaim.lmdb` keys are 1-based, as printed by ORCA and Multiwfn, and follow the atom order of `structure.lmdb`; per-atom entries are keyed `"{i}_{El}"`. `bond.lmdb` keys have the form `"{i}_{El_i}_to_{j}_{El_j}"` and

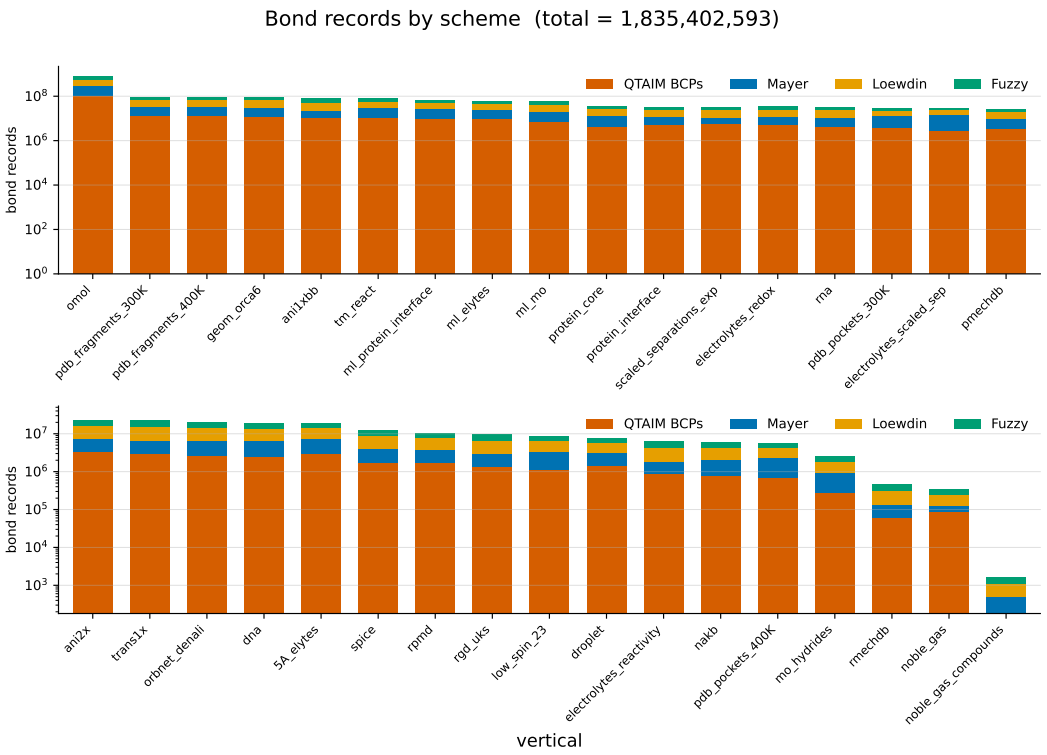

Figure 2: Per-vertical bond-order record counts, stacked by scheme (QTAIM, Mayer, Löwdin, fuzzy) on a log axis. Verticals are sorted by total bond-record count. The schemes report different atom-pair sets (different print thresholds, and QTAIM only where a bond critical point is found; §3), so a bonded pair can appear under up to four schemes and counts represent records, not unique bond instances. Corpus total across the four schemes is 1.83 billion records.

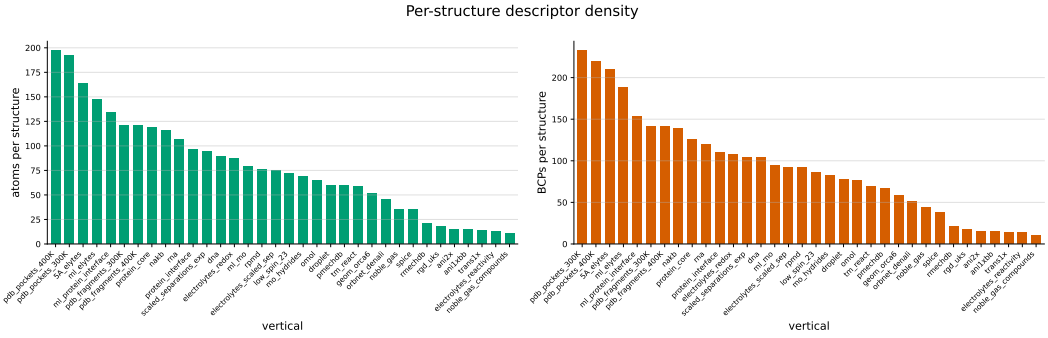

Figure 3: Per-structure descriptor density across the 34 verticals. Left: atoms per structure. Right: QTAIM bond critical points per structure. Solvated-protein verticals (`pdb_pockets_300K`, `pdb_pockets_400K`, `protein_core`) form the high-density tail; `noble_gas_compounds` sits at the low end as the sparse limit.

each pair is stored once, with $i < j$; pairs should be read as undirected. The `qtaim_embed` reader (`parse_bond_data`) converts keys to 0-based sorted tuples. Each bond-order scheme lists only pairs that pass its print threshold (Mayer above 0.10 and Löwdin above 0.05 in ORCA; total fuzzy bond order of at least 0.05 in Multiwfn), so a pair absent from a scheme was below that threshold, not missing. QTAIM bonds are the atom pairs of the bond critical points in `qtaim.lmdb`.

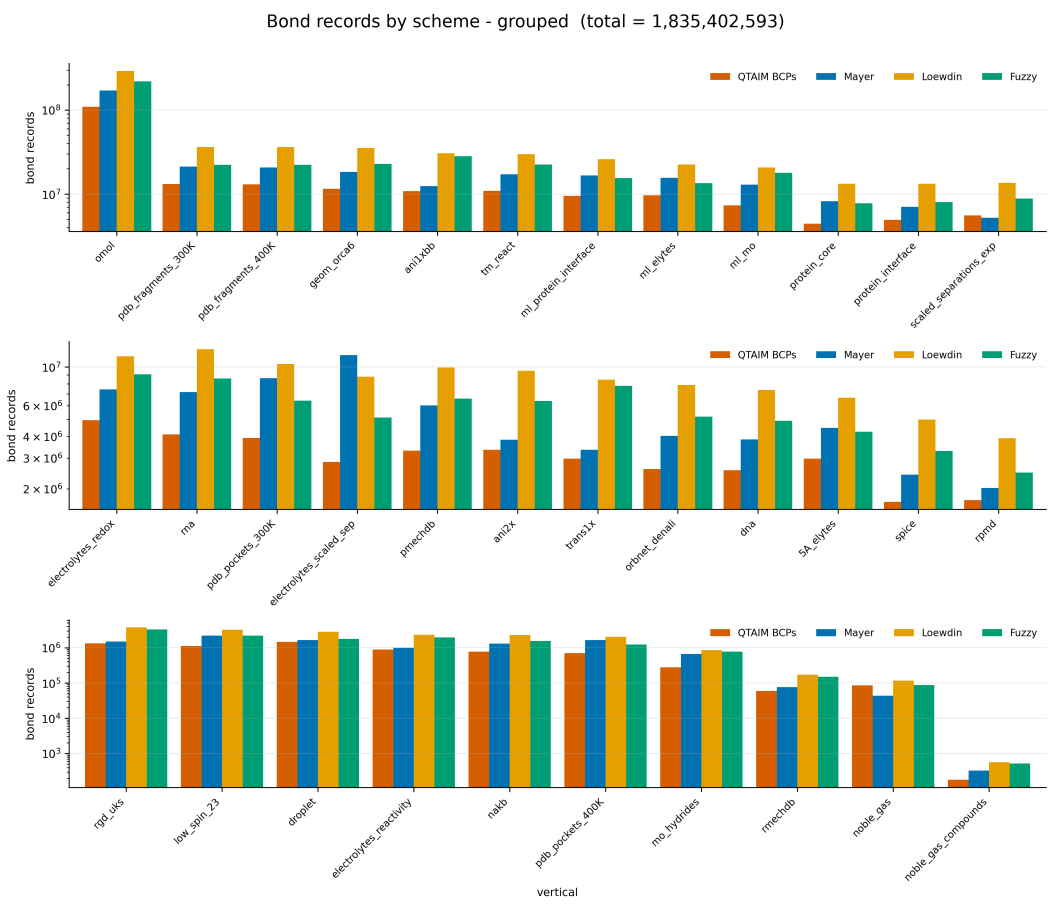

Figure 4: Per-vertical bond-record counts shown as grouped bars rather than stacked. The log-scale stack in Fig. 2 compresses the non-QTAIM schemes against the QTAIM BCP totals; the grouped view confirms that Mayer, Löwdin, and fuzzy bond records are within roughly an order of magnitude of QTAIM BCP counts on every vertical, i.e. the four schemes have comparable coverage rather than QTAIM dominating.

F Per-Vertical Post-Split Distribution

Table 6 reports the realized train / val / test counts per vertical after held-out exclusion, referenced from §5.2. Because the split is global by composition rather than stratified per vertical, realized per-vertical ratios deviate from the nominal $0.8/0.1/0.1$; verticals holding few unique compositions relative to their row count deviate most, since one composition carries many rows there. Row counts are exact: for every vertical, total = train + val + test + held-out, and the column totals recover the full 3,986,738-structure corpus.

G Split Leakage Audit

Composition hashing removes exact-composition leakage but does not by itself control scaffold- or homologue-level similarity, so we quantify both. No test structure shares its exact composition with any training structure (0.0 %), against 80.9 % for a row-level random split of the same corpus: same-composition and conformer leakage are eliminated rather than reduced. Residual similarity is real but bounded. Taking each test composition's nearest training neighbor among compositions with an identical element set, 96.0 % of test compositions have such a neighbor and it differs by a median of 5.3 % of the atom count. For calibration, withholding the `tm_react` vertical entirely

Table 6: Per-vertical post-split distribution. *Held-out* counts are structures removed into the five stress suites before hashing; a structure in two suites is counted once. Train/val/test are the realized as-shipped assignment.

Vertical	Total	Held-out	Train	Val	Test
omol	1,429,192	14,212	1,115,172	152,175	147,633
ani1xbb	737,210	0	586,925	74,952	75,333
ani2x	215,642	0	168,783	17,417	29,442
trans1x	213,731	0	175,394	17,893	20,444
geom_orca6	199,805	0	159,516	20,307	19,982
tm_react	158,190	9,810	119,020	14,634	14,726
pdb_fragments_300K	93,334	30	74,823	9,236	9,245
pdb_fragments_400K	92,698	27	74,253	9,226	9,192
ml_mo	77,603	2,041	60,736	7,554	7,272
rgd_uks	75,316	0	59,417	8,395	7,504
electrolytes_reactivity	66,974	5,073	49,954	5,588	6,359
ml_protein_interface	61,859	1,026	48,838	6,157	5,838
scaled_separations_exp	53,219	2,025	41,003	5,232	4,959
ml_elytes	51,599	7,678	35,092	4,434	4,395
orbnet_denali	51,102	7	40,692	5,199	5,204
pmechdb	49,832	2,506	37,872	4,795	4,659
electrolytes_redox	46,081	1,600	35,516	4,602	4,363
protein_interface	44,972	3	36,257	4,511	4,201
spice	44,450	3	35,335	4,415	4,697
protein_core	35,406	129	28,429	3,715	3,133
rna	34,745	2,538	25,675	4,173	2,359
electrolytes_scaled_sep	30,769	555	24,178	3,111	2,925
dna	24,647	190	21,119	2,005	1,333
droplet	18,694	26	14,933	1,951	1,784
rpmd	18,652	0	15,068	1,808	1,776
pdb_pockets_300K	16,927	4,820	9,689	1,150	1,268
5A_elytes	14,202	2,441	9,299	1,224	1,238
low_spin_23	12,939	125	10,288	1,269	1,257
nakb	5,606	768	3,797	771	270
mo_hydrides	3,396	160	2,602	322	312
pdb_pockets_400K	3,210	993	1,778	242	197
rmechdb	2,803	0	2,090	343	370
noble_gas	1,915	0	1,462	173	280
noble_gas_compounds	18	0	14	3	1
Total	3,986,738	58,786	3,125,019	398,982	403,951

raises that median to 16.5 %, so the composition split sits roughly one third of the way from a random split to a full vertical holdout. A direct homologue probe finds that 17.3 % of test compositions have a training composition differing by exactly one CH_2 unit. We report these measurements rather than adding a Murcko scaffold split because scaffold decomposition requires a sanitized covalent bond graph, which is undefined for the transition-metal, ionic, multi-fragment and reaction-path verticals that motivate this corpus: 27.0 % of structures contain an element outside the organic subset that scaffold perception assumes, and 19.0 % contain a transition metal, lanthanide or actinide. A further observation supports the stress-suite design in §5.3 over per-vertical holdout: withholding `droplet` or `rna` entirely still leaves 67–68 % of their compositions matched exactly in training, because solvent clusters and nucleotide fragments recur across verticals.

H LMDB Conversion Quality Audit

Table 7 reports raw failure counts per vertical from the post-conversion audit (`lmdb_status_audit.py`). *QTAIM fail*: the `qtaim.lmdb` record holds no nuclear critical point with atomic properties. *ORCA fail*: the `orca.lmdb` record lacks the final SCF energy (includes SCF non-convergence flagged at the parse stage). *Other fail*: failures across `structure`,

charge, bond, fuzzy, and other LMDBs (all zero in this release). *Max drop*: largest number of structure.lmdb keys absent from any single descriptor LMDB; values ≤ 4 across all verticals.

Table 7: Per-vertical LMDB conversion-quality audit, run on the converted LMDBs before release filtering. Verticals with all zeros across the four failure columns are shown as "–". Jobs counted as QTAIM or ORCA failures are removed from every descriptor LMDB in the release (not left as partial records): 1,631 jobs (1,632 failure counts; one ml_mo job fails both), which is the gap between the 3,986,738 catalogued structures and the 3,985,107 carried by the descriptor LMDBs.

Vertical	N	QTAIM fail	ORCA fail	Other fail	Max drop
omol	1,429,192	17	896	0	1
ani1xbb	737,210	2	0	0	0
ani2x	215,642	–	–	–	–
trans1x	213,731	–	–	–	–
geom_orca6	199,805	–	–	–	–
tm_react	158,190	0	1	0	0
pdb_fragments_300K	93,334	0	0	0	1
pdb_fragments_400K	92,698	–	–	–	–
ml_mo	77,603	1	23	0	0
rgd_uks	75,316	–	–	–	–
electrolytes_reactivity	66,974	0	7	0	0
ml_protein_interface	61,859	0	3	0	0
scaled_separations_exp	53,219	2	37	0	0
ml_elytes	51,599	1	38	0	0
orbnet_denali	51,102	–	–	–	–
pmechdb	49,832	–	–	–	–
electrolytes_redox	46,081	0	12	0	0
protein_interface	44,972	0	212	0	0
spice	44,450	1	2	0	4
protein_core	35,406	–	–	–	–
rna	34,745	0	311	0	0
electrolytes_scaled_sep	30,769	0	1	0	0
dna	24,647	–	–	–	–
droplet	18,694	–	–	–	–
rpmd	18,652	0	0	0	1
pdb_pockets_300K	16,927	0	57	0	0
5A_elytes	14,202	0	2	0	2
low_spin_23	12,939	–	–	–	–
nakb	5,606	–	–	–	–
mo_hydrides	3,396	–	–	–	–
pdb_pockets_400K	3,210	1	5	0	0
rmechdb	2,803	–	–	–	–
noble_gas	1,915	–	–	–	–
noble_gas_compounds	18	–	–	–	–
Total	3,986,738	25	1607	0	4

Table 8 reports the released record count of each LMDB after the failed jobs are removed. Relative to structure.lmdb, 2 keys are absent from the charge, bond, fuzzy, other and timings LMDBs, 3 from orca.lmdb and 7 from qtaim.lmdb; energy.lmdb carries every catalogued structure, including the 1,631 failed jobs.

Figure 5 shows per-method descriptor coverage across all 34 verticals. Rows are descriptor methods grouped by family; columns are verticals ordered by structure count (largest left). Universal methods (Hirshfeld, ADCH, CM5, Becke charges; Fuzzy bond orders; ORCA Mayer/Löwdin/SCF/gradient/dipole) reach $\geq$99% on every vertical. QTAIM critical-point records achieve $\geq$99.9% coverage.

Table 8: Per-LMDB record counts in the release, by split. *Held-out* counts distinct structures across H1–H8 (58,786; a structure in two suites is counted once).

LMDB	Train	Val	Test	Held-out	Total
`structure`	3,123,738	398,813	403,770	58,786	3,985,107
`charge`, `bond`, `fuzzy`, `other`, `timings`	3,123,737	398,813	403,770	58,785	3,985,105
`orca`	3,123,737	398,812	403,770	58,785	3,985,104
`qtaim`	3,123,733	398,813	403,769	58,785	3,985,100
`energy`	3,125,019	398,982	403,951	58,786	3,986,738

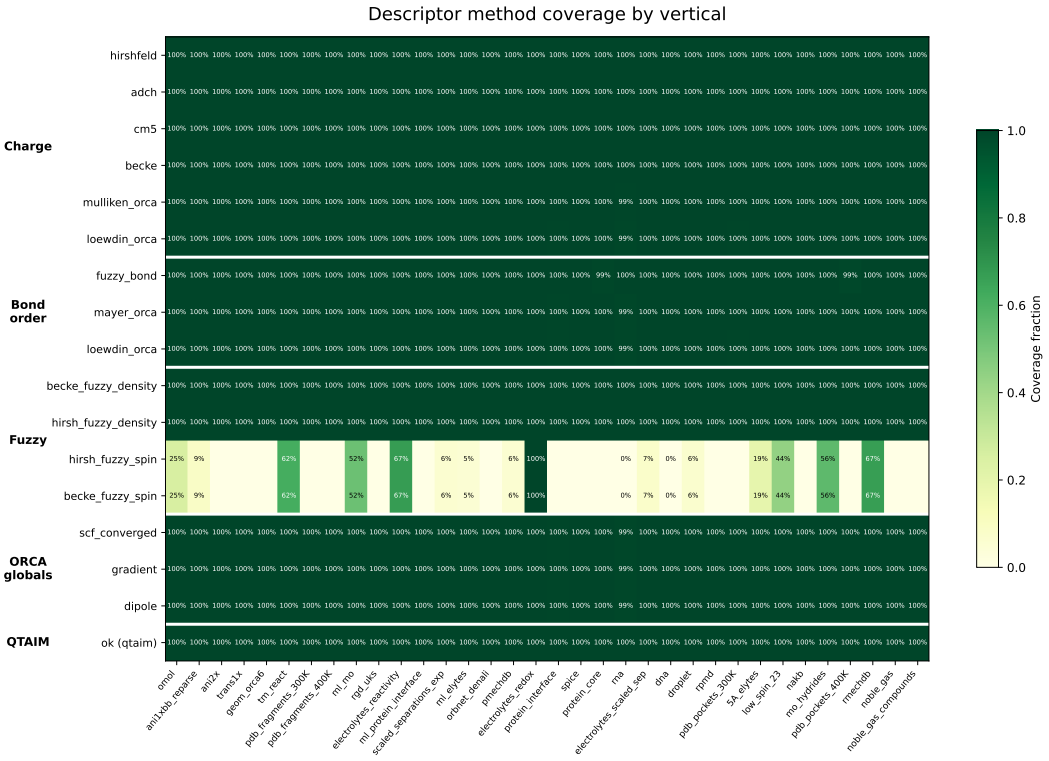

Figure 5: Per-method descriptor coverage fraction across all 34 verticals. Green = full coverage; white = absent. Rows are methods and columns are verticals; horizontal white lines separate descriptor families. The bottom `ok (qtaim)` row reports `qtaim.lmdb` record validity. The fuzzy spin rows are defined only for open-shell records, so their coverage equals each vertical's open-shell fraction rather than missing data. The `ani1xbb_reparse` label is the `ani1xbb` vertical.

I Energy and Force Labels

`energy.lmdb` is a key-space join with the OMol25 release, not a recomputation. Each record holds the OMol25 total energy and forces with atoms in the same order as `structure.lmdb`, the OMol25 record identifier, and OMol25's per-record metadata (charge, spin, electron and ECP-electron counts, $\langle S^2 \rangle$, HOMO energy and gap, and OMol25's Mulliken and Löwdin charges). It covers all 3,986,738 catalogued structures, including the jobs absent from the descriptor LMDBs (§3.3). Against the final SCF energies and gradients we parse independently into `orca.lmdb`, every shared record in every split and held-out suite agrees to within 2.9×10^{-11} eV in energy and 7.1×10^{-15} eV/Å in force, provided Hartree is converted with ASE's constant (CODATA 2014), as OMol25 does; CODATA 2018 shifts the largest systems by up to 0.023 eV.

J CHARGE-SCHEME CROSS-CORRELATIONS

We quantify pairwise agreement across the six partial-charge schemes (Hirshfeld, CM5, ADCH, Becke, Mulliken, Löwdin) using two symmetric metrics computed over all atom-level charge records in the corpus: pairwise R^2 (coefficient of determination between scheme a and scheme b over all atoms) and mean absolute error (MAE, in electron units). Figures 6 and 7 show the corpus-wide matrices; Figures 8 and 9 break these down by element class and by individual transition metal; Figure 10 shows per-vertical matrices.

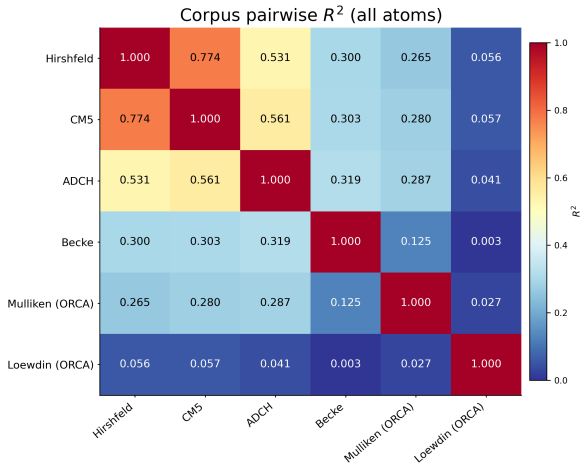

Figure 6: Corpus-wide pairwise R^2 between charge schemes over all ∼218M atom records. Hirshfeld–CM5 ($R^2 = 0.77$) and Hirshfeld–ADCH ($R^2 = 0.53$) show the strongest cross-scheme agreement; Mulliken and Löwdin correlate weakly with the stockholder-partitioned schemes ($R^2 \leq 0.30$).

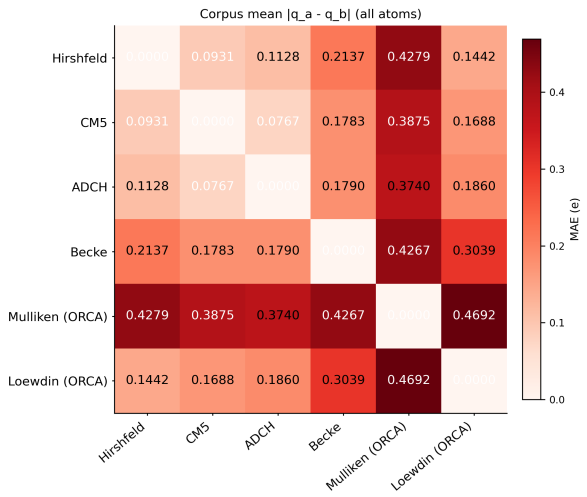

Figure 7: Corpus-wide pairwise MAE (in e) between charge schemes. CM5–ADCH (0.077 e) has the smallest cross-scheme deviation, followed by Hirshfeld–CM5 (0.093 e) and Hirshfeld–ADCH (0.113 e). Mulliken diverges most ($\geq$0.37 e versus Hirshfeld/CM5/ADCH; 0.469 e versus Löwdin), consistent with its basis-set sensitivity at the def2-TZVPD level used throughout OMol25.

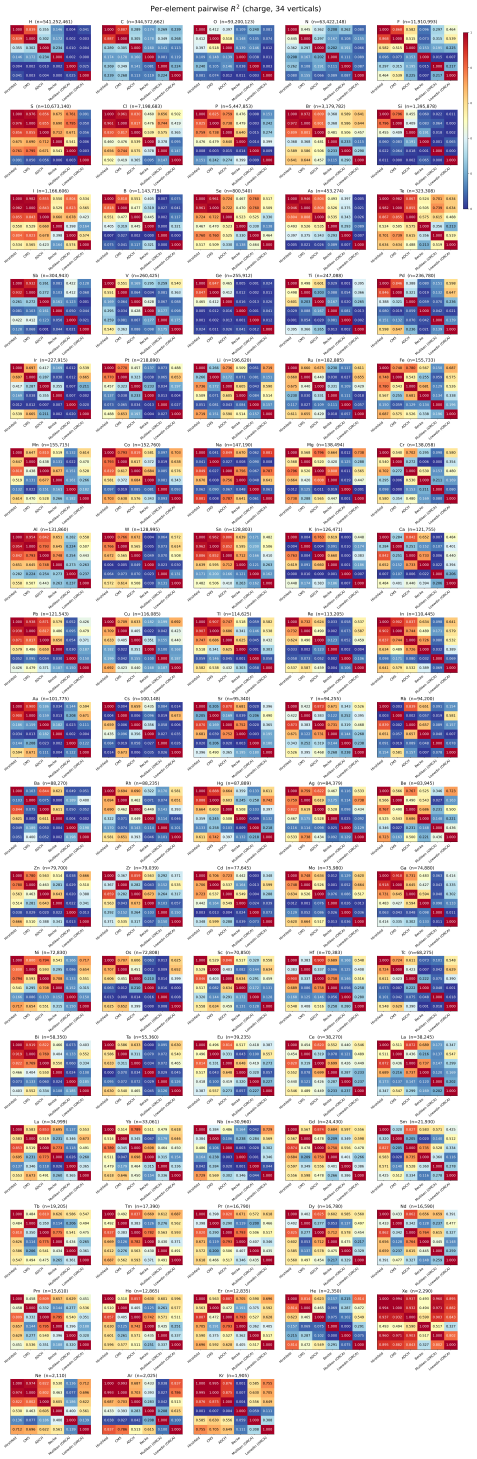

Figure 8: Pairwise R^2 per element (one panel per element), with panels grouped by element class: main-group, transition-metal, lanthanide, and noble-gas. Inter-scheme agreement tracks the corpus-wide pattern within main-group chemistry; transition metals and lanthanides show reduced Hirshfeld–ADCH agreement, reflecting the influence of d/f polarization on atomic-basin integration.

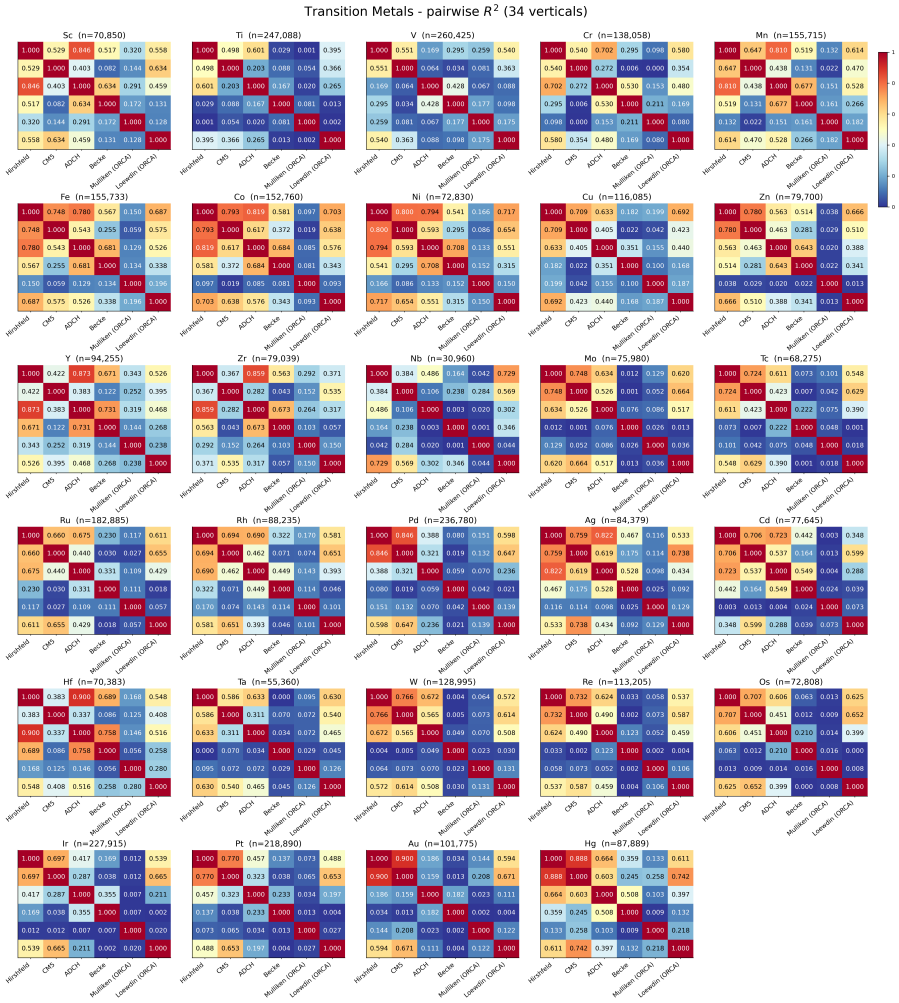

Figure 9: Pairwise R^2 broken down by individual transition-metal element (30 panels, n = atoms bearing that TM across all 34 verticals). Each panel uses the same 6-scheme axis as Figure 6. Variation across TMs is substantial: early $3d$ metals (Sc, Ti, V) retain moderate Hirshfeld–CM5 agreement ($R^2 \sim 0.4$–0.5) while late $4d$/$5d$ metals (Pd, Pt, Au) drop to $R^2 < 0.3$, indicating that the connectivity correction in CM5 diverges from the stockholder partition for electron-rich coordination environments.

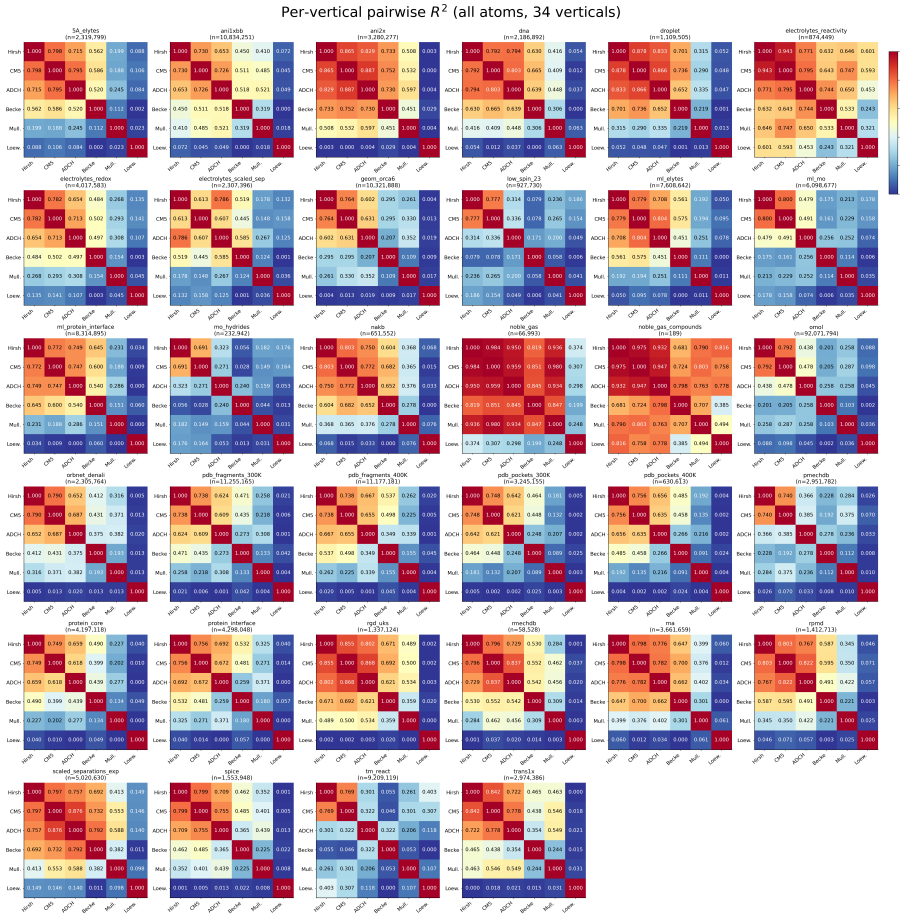

Figure 10: Per-vertical pairwise R^2 across all 34 verticals (34 panels). Organic-heavy verticals (`ani1xbb`, `spice`, `trans1x`) reproduce the high Hirshfeld–CM5 corpus agreement; TM-rich verticals (`tm_react`, `ml_mo`, `mo_hydrides`) show characteristically lower cross-scheme R^2, particularly for Becke and Löwdin. `noble_gas` and `noble_gas_compounds` have near-zero off-diagonal entries, as partial charges on noble-gas atoms are close to zero for all schemes.

K CROSS-SCHEME NOISE FLOORS

Two descriptor families are computed by more than one scheme on the same structures: six atomic partial charges (Hirshfeld, CM5, ADCH and Becke from Multiwfn; Mulliken and Löwdin from ORCA) and three bond orders (fuzzy from Multiwfn; Mayer and Löwdin from ORCA). Every scheme is evaluated on the identical atom or atom pair, so the spread between schemes reflects the definitions themselves rather than sampling. We summarize it as the *cross-scheme median absolute residual* (MAR): for schemes a and b, MAR_{ab} is the median of $|x_a - x_b|$ over all atoms (charges, in e) or over all atom pairs reported by both schemes (bond orders, in bond-order units). The median keeps the statistic insensitive to the heavy tails of ECP-treated metal charges. The per-vertical *noise floor* is the median of MAR_{ab} over the 15 charge-scheme pairs, or over the 3 bond-order pairs. It is a reference scale for model errors on these targets, not an estimate of the error of any one scheme.

Table 9 lists the floors for all 34 verticals (218.5M atoms for charges; 378.4M closed-shell atom pairs for bond orders). Charge floors span 0.068–0.207 e (median 0.096 e), lowest for the electrolyte and protein verticals (`ml_elytes` 0.068 e, `protein_core` 0.069 e) and highest for nucleic acids, droplets and noble-gas systems (`nakb` 0.207, `rna` 0.202, `noble_gas` 0.183, `dna` 0.179, `droplet` 0.177 e). Closed-shell bond-order floors span 0.116–0.303 (median 0.214).

Per scheme (Figure 11), Mulliken departs most from the other charge schemes in 31 of 34 verticals, and Löwdin in the remaining three (`droplet`, `noble_gas`, `noble_gas_compounds`). The five largest pairwise charge MARs all involve Mulliken (corpus medians 0.26–0.34 e); every other pair is at or below 0.14 e, with CM5–ADCH the closest pair (0.029 e). Because the floor is a median over 15 pairs, it is set by the non-Mulliken pairs in every vertical. Becke here denotes the ADC-corrected Becke charges shipped in `charge.lmdb` (App. M).

Open-shell bond orders. Bond-order floors use closed-shell records only. `electrolytes_redox` has no closed-shell records; for it we report the Mayer–Löwdin MAR alone, which uses ORCA bond orders only.

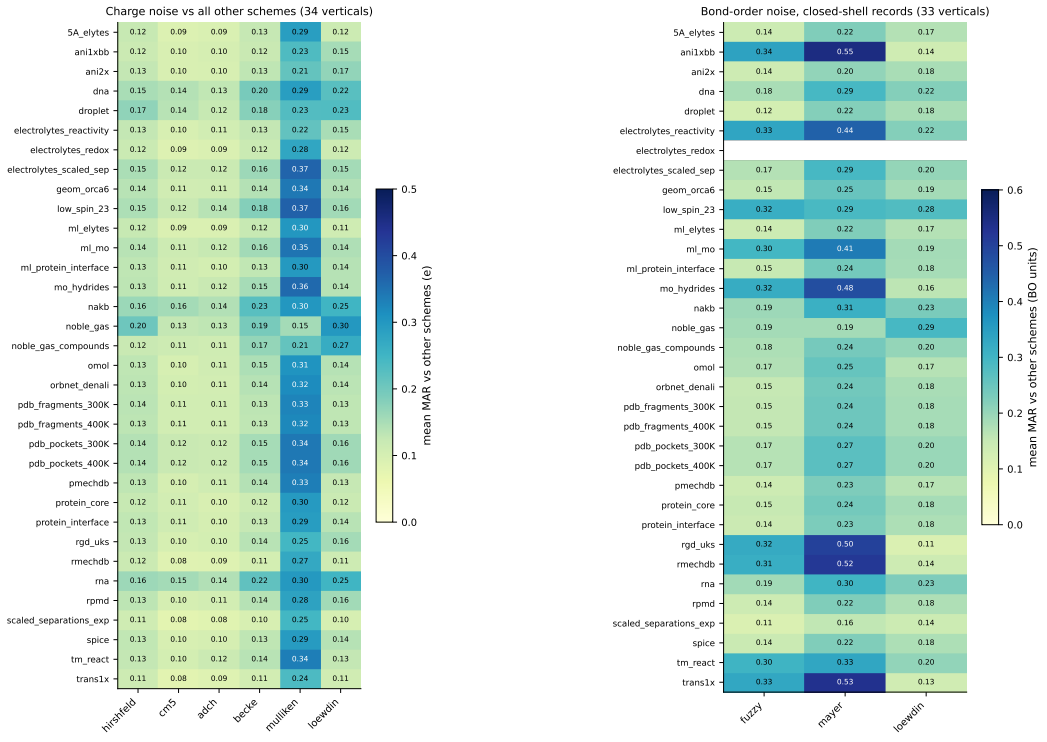

Figure 11: Per-scheme disagreement: for each vertical and scheme, the n-weighted mean MAR against every other scheme of the same family. Left: charges (all 34 verticals, e). Right: bond orders on closed-shell records (33 verticals; `electrolytes_redox` has none), bond-order units.

Table 9: Per-vertical cross-scheme noise floors: median over scheme pairs of the median absolute residual between schemes on identical atoms (charges, 15 pairs) or atom pairs (bond orders, 3 pairs), with the number of atoms and closed-shell atom pairs. †`electrolytes_redox` has no closed-shell records; its entry is the Mayer–Löwdin MAR over all records.

Vertical	Charge floor (e)	Bond floor (BO)	Atoms	Atom pairs
`5A_elytes`	0.071	0.192	2,324,023	3,350,733
`ani1xbb`	0.113	0.224	10,834,251	24,533,425
`ani2x`	0.137	0.146	3,280,277	6,375,147
`dna`	0.179	0.249	2,186,892	4,903,478
`droplet`	0.177	0.153	1,109,505	1,687,743
`electrolytes_reactivity`	0.112	0.269	874,449	681,195
`electrolytes_redox`	0.071	0.240†	4,017,583	8,778,537†
`electrolytes_scaled_sep`	0.098	0.274	2,307,396	5,140,242
`geom_orca6`	0.086	0.216	10,321,888	22,826,559
`low_spin_23`	0.113	0.303	927,730	1,206,501
`ml_elytes`	0.068	0.194	7,608,642	12,703,288
`ml_mo`	0.096	0.261	6,098,677	8,503,742
`ml_protein_interface`	0.095	0.204	8,314,895	15,474,467
`mo_hydrides`	0.090	0.270	232,942	342,138
`nakb`	0.207	0.271	651,552	1,550,275
`noble_gas`	0.183	0.266	66,993	86,067
`noble_gas_compounds`	0.171	0.200	189	518
`omol`	0.092	0.244	92,071,794	148,404,584
`orbnet_denali`	0.082	0.205	2,305,764	5,186,714
`pdb_fragments_300K`	0.076	0.214	11,255,165	22,277,336
`pdb_fragments_400K`	0.076	0.212	11,177,181	22,202,329
`pdb_pockets_300K`	0.102	0.242	3,245,155	6,393,120
`pdb_pockets_400K`	0.106	0.241	630,613	1,219,131
`pmechdb`	0.082	0.197	2,951,782	6,092,683
`protein_core`	0.069	0.204	4,197,118	7,734,264
`protein_interface`	0.111	0.192	4,298,048	8,001,492
`rgd_uks`	0.136	0.172	1,337,124	3,170,011
`rmechdb`	0.079	0.254	58,528	52,880
`rna`	0.202	0.262	3,661,659	8,577,347
`rpmd`	0.102	0.179	1,412,713	2,471,467
`scaled_separations_exp`	0.069	0.116	5,020,630	8,300,777
`spice`	0.085	0.183	1,553,948	3,290,995
`tm_react`	0.085	0.208	9,209,119	8,420,137
`trans1x`	0.075	0.219	2,974,386	7,232,911

L T1 Results by Charge Scheme

Per-scheme breakdown of Table 3: the same models, splits and structure counts (§5.5). MAE is per atom in e; $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$ over all atoms of a set. The best model in each row is in bold.

Table 10: T1 MAE (e) by charge scheme, test split and held-out sets.

Set	Scheme	GCN	GAT	SchNet	DimeNet++	Equivariant
Test	ADCH	0.0570	0.0652	0.0508	0.0479	**0.0416**
	Becke	0.1304	0.1435	0.1114	0.0918	**0.0853**
	CM5	0.0101	0.0165	0.0096	0.0112	**0.0092**
	Hirshfeld	0.0096	0.0161	0.0077	0.0108	**0.0072**
	Löwdin	0.0140	0.0222	0.0109	0.0129	**0.0105**
	Mulliken	0.1487	0.1821	0.0899	0.0901	**0.0801**
H1	ADCH	0.0898	0.1043	0.0806	0.0736	**0.0631**
	Becke	0.2099	0.2326	0.1803	0.1472	**0.1361**
	CM5	0.0100	0.0196	0.0098	0.0115	**0.0093**
	Hirshfeld	0.0090	0.0185	0.0083	0.0112	**0.0078**
	Löwdin	0.0150	0.0259	0.0120	0.0133	**0.0117**
	Mulliken	0.1563	0.2148	0.1136	0.1196	**0.1038**
H3	ADCH	0.0743	0.0889	0.0705	0.0716	**0.0565**
	Becke	0.1312	0.1519	0.1196	0.1115	**0.0916**
	CM5	0.0153	0.0282	0.0153	0.0199	**0.0137**
	Hirshfeld	0.0147	0.0272	0.0144	0.0192	**0.0127**
	Löwdin	0.0191	0.0331	0.0169	0.0207	**0.0155**
	Mulliken	0.1236	0.1763	0.0946	0.1004	**0.0837**
H6	ADCH	0.0818	0.0940	0.0757	0.0690	**0.0588**
	Becke	0.1794	0.2000	0.1605	0.1337	**0.1194**
	CM5	0.0103	0.0208	0.0106	0.0120	**0.0098**
	Hirshfeld	0.0094	0.0192	0.0088	0.0110	**0.0080**
	Löwdin	0.0153	0.0280	0.0127	0.0140	**0.0122**
	Mulliken	0.1300	0.1829	0.0939	0.0986	**0.0852**
H7	ADCH	0.0327	0.0360	0.0266	0.0239	**0.0237**
	Becke	0.0802	0.0872	0.0636	**0.0498**	0.0502
	CM5	0.0086	0.0103	**0.0076**	0.0079	0.0078
	Hirshfeld	0.0084	0.0107	**0.0055**	0.0068	0.0057
	Löwdin	0.0127	0.0158	**0.0084**	0.0104	0.0084
	Mulliken	0.3140	0.3236	0.1368	0.1302	**0.1212**
H8	ADCH	0.0552	0.0624	0.0462	0.0423	**0.0396**
	Becke	0.0882	0.0975	0.0763	0.0644	**0.0609**
	CM5	0.0133	0.0187	**0.0111**	0.0112	0.0115
	Hirshfeld	0.0134	0.0188	**0.0092**	0.0102	0.0098
	Löwdin	0.0164	0.0233	**0.0115**	0.0129	0.0118
	Mulliken	0.2003	0.2371	0.1068	0.1082	**0.0959**

Table 11: T1 R^2 by charge scheme, test split and held-out sets.

Set	Scheme	GCN	GAT	SchNet	DimeNet++	Equivariant
Test	ADCH	0.656	0.613	0.693	0.759	**0.789**
	Becke	0.452	0.411	0.542	**0.735**	0.727
	CM5	0.991	0.976	0.991	0.985	**0.992**
	Hirshfeld	0.971	0.929	0.975	0.951	**0.978**
	Löwdin	0.991	0.972	0.994	0.989	**0.994**
	Mulliken	0.816	0.749	0.930	0.931	**0.944**
H1	ADCH	0.531	0.444	0.601	0.705	**0.760**
	Becke	0.280	0.209	0.435	0.674	**0.679**
	CM5	0.994	0.973	0.994	0.990	**0.994**
	Hirshfeld	0.982	0.921	0.983	0.965	**0.985**
	Löwdin	0.993	0.976	0.995	0.993	**0.995**
	Mulliken	0.860	0.757	0.919	0.912	**0.932**
H3	ADCH	0.556	0.477	0.578	0.633	**0.735**
	Becke	0.382	0.301	0.446	0.579	**0.675**
	CM5	0.980	0.927	0.980	0.959	**0.984**
	Hirshfeld	0.960	0.853	0.960	0.915	**0.968**
	Löwdin	0.984	0.942	0.986	0.975	**0.988**
	Mulliken	0.894	0.792	0.935	0.925	**0.948**
H6	ADCH	0.529	0.467	0.576	0.677	**0.738**
	Becke	0.283	0.232	0.385	0.615	**0.660**
	CM5	**0.995**	0.978	0.994	0.992	0.995
	Hirshfeld	0.985	0.937	0.985	0.975	**0.988**
	Löwdin	0.993	0.971	0.995	0.993	**0.995**
	Mulliken	0.894	0.798	0.941	0.937	**0.951**
H7	ADCH	0.805	0.795	0.821	**0.886**	0.859
	Becke	0.632	0.616	0.656	**0.862**	0.803
	CM5	0.983	0.981	**0.984**	0.984	0.981
	Hirshfeld	0.936	0.925	**0.943**	0.941	0.935
	Löwdin	0.989	0.979	**0.994**	0.992	0.992
	Mulliken	0.551	0.509	0.896	0.908	**0.917**
H8	ADCH	0.679	0.675	0.711	0.784	**0.815**
	Becke	0.637	0.626	0.667	0.781	**0.810**
	CM5	0.977	0.965	**0.979**	0.979	0.975
	Hirshfeld	0.945	0.922	**0.953**	0.952	0.944
	Löwdin	0.980	0.949	**0.989**	0.986	0.987
	Mulliken	0.711	0.641	0.909	0.912	**0.926**

M CROSS-PACKAGE VALIDATION

We re-derived a subset of the released descriptors with two codes independent of Multiwfn, reading the same wavefunction files used in production. The subset holds 87 structures from seven verticals chosen to span the regimes where partitioning is easy and where it is hard: organics (`rmechdb`), transition-metal and lanthanide complexes (`tm_react`, `mo_hydrides`), charged solvation shells (`5A_elytes`, `droplet`), and noble-gas systems (`noble_gas`, `noble_gas_compounds`); 25 of the 87 use ECPs.

Partial charges (HORTON). Becke and Hirshfeld charges were recomputed with HORTON (Chan et al., 2024) on molecular integration grids from `qc-grid` (Tehrani et al., 2024). Hirshfeld was run on the 62 all-electron structures only, since the pro-atom densities are all-electron. The two Hirshfeld implementations also use different pro-atom libraries (Multiwfn's built-in spherical atomic densities versus `qc-AtomDB`), so their residual bounds implementation and reference-density effects together. Becke partitions depend on an atomic-radii table and a clip on the size-adjustment parameter; HORTON defaults to Bragg–Slater radii and Multiwfn to modified CSD radii. With native conventions the two codes correlate at only $r = 0.49$; with Multiwfn's radii and clip reproduced in HORTON they agree as closely as Hirshfeld does (Table 12). The radii table is therefore a free parameter of the Becke scheme, which we disclose rather than tune. The released `becke` scheme is additionally ADC-corrected (Sec. 3), which is why raw HORTON Becke does not track it. HORTON has no ADC step, so ADC-Becke and ADCH remain single-implementation, as do CM5 and the fuzzy bond orders. The largest Hirshfeld residuals (0.05–0.06 e) all fall on metal centers (Ga, Ca, Mn, Zn, Ni, Al).

Table 12: HORTON versus Multiwfn atomic charges on the same wavefunctions. "Matched radii" uses Multiwfn's modified-CSD radii and ± 0.5 size-adjustment clip in HORTON. Differences in e.

Comparison	Atoms	Median $\lvert\Delta q\rvert$	Max $\lvert\Delta q\rvert$	Pearson r
Hirshfeld, HORTON vs. Multiwfn	1,030	0.004	0.062	0.9993
Raw Becke, matched radii	1,355	0.005	0.848	0.9940
Raw Becke, native radii	1,355	0.233	2.735	0.489
Raw Becke vs. released ADC-Becke	1,355	0.259	5.725	0.043

QTAIM topology (Critic2). Critical points were located with Critic2 (Otero-de-la Roza et al., 2014) on the same wavefunctions and matched to the Multiwfn bond critical points by atom pair. Table 13 reports agreement on the 1,439 BCPs found by both codes. Median differences are below 10^{-6} a.u. for every field except ELF and LOL, which the two codes define against different uniform-electron-gas references and so differ by a systematic offset. On the 62 all-electron structures every other field agrees with $r = 1.0000$. The BCP sets themselves, which are the graph edges under the QTAIM bonding scheme, match exactly in 77 of 87 structures (59 of 62 all-electron, 18 of 25 ECP); 10 BCPs are found only by Critic2 and 29 only by Multiwfn, 31 of these 39 in ECP structures and 24 at $\rho < 0.05$ a.u. The two codes sometimes differ on individual BCPs: about 1% of shared BCPs disagree in ellipticity (> 0.1) or η (> 0.01), all in ECP structures and concentrated at low density (median $\rho \approx 0.05$ a.u. versus 0.26 a.u. for agreeing BCPs), where the two smaller Hessian eigenvalues are nearly degenerate and their ratio is ill-conditioned. Critic2 does not read the EDF core densities, so near some ECP atoms (10 structures) it reports non-nuclear attractors, which we assign to the nearest nucleus.

N SOAP-UMAP STRUCTURAL OVERLAP WITH COMPARATOR DATASETS

Figure 12 places OMol-Descriptors-4M against the two organic descriptor comparators (QM7-X, SchNet4AIM). The UMAP is fit on a balanced anchor of 500 structures from each of five sources (2,500 points), including PCQM4Mv2 and QMugs, which are not plotted.

That comparison uses a 12-element species set (H, B, C, N, O, F, Si, P, S, Cl, Br, I) shared with the organic comparators. That basis silently drops transition-metal atoms during the SOAP fingerprint and therefore cannot fairly place tmQM+ (Gee et al., 2025) on the same map. We refit a separate UMAP at a 32-element TM-inclusive species basis – the 12 organic species plus the first-row transition metals (Sc–Zn) and the 2nd/3rd-row TMs (Mo, Ru, Rh, Pd, W, Re, Ir, Pt, Au, Hg) that exceed 1% structure

Table 13: Critic2 versus Multiwfn on 1,439 shared BCPs (1,438 for Hessian-derived fields). Atomic units. †Convention-divergent definition.

Field	Median $\lvert\Delta\rvert$	Mean $\lvert\Delta\rvert$	r (87 structures)	r (62 all-electron)
ρ	1.9×10^{-8}	1.7×10^{-4}	0.9992	1.0000
$\nabla^2\rho$	8.7×10^{-8}	8.9×10^{-4}	0.9995	1.0000
Lagrangian kinetic energy G	5.3×10^{-9}	7.4×10^{-5}	1.0000	1.0000
Hamiltonian kinetic energy K	1.8×10^{-8}	2.7×10^{-4}	0.9994	1.0000
Energy density H	1.8×10^{-8}	3.9×10^{-4}	0.9985	1.0000
Hessian determinant	5.5×10^{-8}	1.7×10^{-2}	0.9471	1.0000
η	3.1×10^{-7}	1.5×10^{-3}	0.9995	1.0000
Ellipticity	2.8×10^{-7}	4.1×10^{-2}	0.4029	1.0000
ELF†	1.2×10^{-2}	6.0×10^{-2}	0.9792	0.9795
LOL†	4.2×10^{-2}	4.9×10^{-2}	0.9888	0.9903
BCP position (Å)	4.1×10^{-7}	3.4×10^{-4}	–	–

share in either dataset – with $r_{\text{cut}} = 5.0$ Å, $n_{\max} = 8$, $l_{\max} = 6$, dim = 230,272. Per-source SOAP parquets total ∼159k OMol-Descriptors-4M structures (drawn across all 34 verticals plus four held-out suites) and 12,147 tmQM+ structures from the high-LOT split of the tmQM+ figshare deposit. Because pyNNDescent is prohibitively slow at this dimensionality, dim reduction first projects every SOAP vector to 256 components via a deterministic GaussianRandomProjection (Johnson–Lindenstrauss; seed=42); UMAP fits on the projected anchor sample and projects the remaining rows via `reducer.transform`. We sample 50% of the rest set per parquet to cap wall time, leaving 81,945 OMol-Descriptors-4M points and 8,574 tmQM+ points on the canvas. The reproducibility script is `notebooks/fit_umap_pca_full.py` (despite its name, it uses the random projection above, not PCA) and the render is `notebooks/render_umap_omol_vs_tmqmplus.py`.

O Pipeline Details

A user selects a calculation tier through a single integer `full_set` $\in \{0, 1, 2\}$ which cascades through the four post-DFT families implemented in `qtaim_gen/source/data/multiwfn.py`. Tiers are strict supersets, so calculations stack from 0 to 2 as descriptor needs grow. Table 14 summarizes per-family additions at each tier; the QTAIM critical-point search and ORCA-derived globals run at every tier.

Table 14: Per-family contents of each calculation tier (`full_set`). Tiers are strict supersets: tier k runs everything in tiers $0, \ldots, k$. Schemes shown in a column are those *added* at that tier. Source: `qtaim_gen/source/data/multiwfn.py`.

Family	**Tier 0 (baseline)**	**Tier 1 (+adds)**	**Tier 2 (+adds)**
Partial charges	Hirshfeld, ADCH, CM5, Becke	VDD, MBIS, CHELPG	Bader (basin-integrated)
Bond orders	Fuzzy bond order	IBSI bond order	Laplacian bond order
Fuzzy density	Becke density, Hirshfeld density, Becke spin, Hirshfeld spin	ELF, MBIS density, MBIS spin	Laplacian ρ, gradient-norm $\lvert\nabla\rho\rvert$
Other (scalar fields)	Planarity (MPP, SDP), ALIE surface	(none)	(none)
Every Tier: full QTAIM critical-point search (atoms, bond CPs, ring CPs, cage CPs; ρ, $\nabla^2\rho$, ellipticity, energy density); ORCA `.out` parsed globals (energies, orbital eigenvalues, dipole, gradient, Mulliken / Löwdin charges, SCF metadata, quality flags).			

Each job folder tracks its own completion state. A `.processing.lock` file with a stale-mtime threshold prevents conflicting writes across distributed workers, and a per-folder `out_files.zip` archives every Multiwfn stream so a folder can be reparsed without recomputing. Validation is

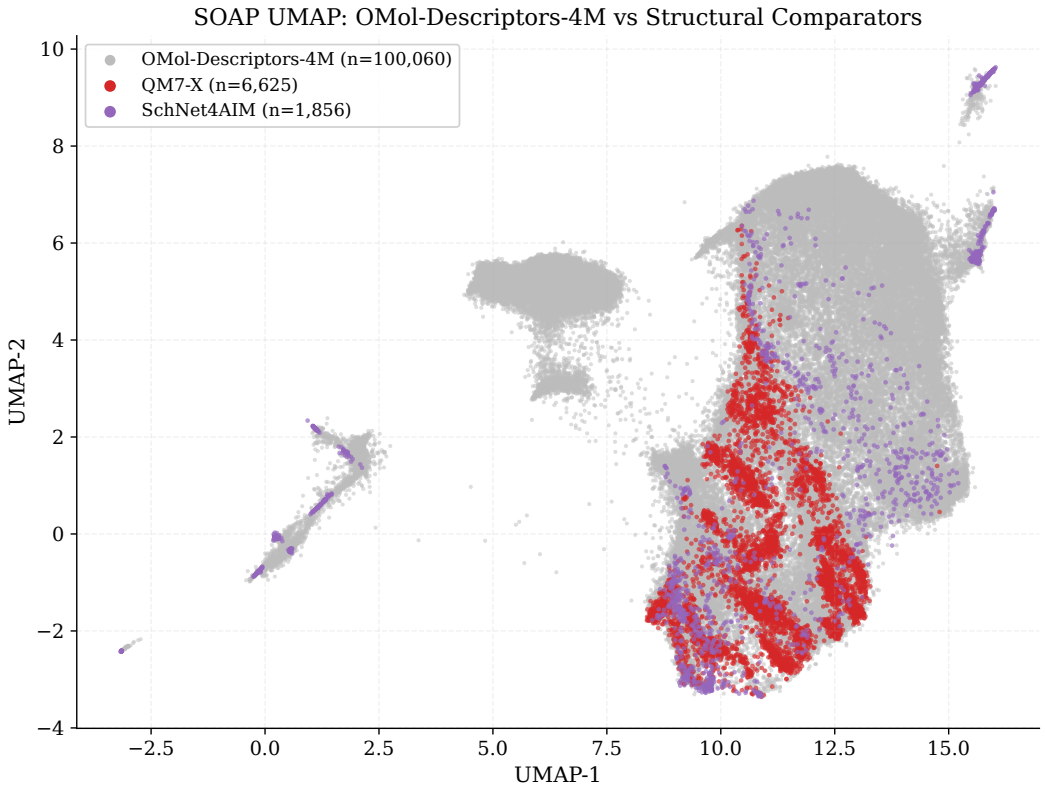

Figure 12: SOAP-UMAP embedding of OMol-Descriptors-4M against the two organic structural comparator datasets. Each point is a structure projected via UMAP fitted on a 500-structure-per-source balanced anchor. SOAP descriptors use a 12-element species set (H, B, C, N, O, F, Si, P, S, Cl, Br, I). Plotted points are the anchor plus a 25 % uniform sample of each source's remaining SOAP records: 6,625 QM7-X (from a 25,000-conformer SOAP sample), 1,856 SchNet4AIM (of 5,925 records) and 100,060 OMol-Descriptors-4M structures. Atoms outside the species set are omitted from the SOAP vector; this affects 27,148 OMol-Descriptors-4M structures (27.1 %, 0.7 % of their atoms) and no comparator structure. The transition-metal-inclusive embedding is in Fig. 13. OMol-Descriptors-4M (gray) is rendered as a background haze to expose the relative positions of QM7-X and SchNet4AIM.

structural: each tier defines the JSON keys that must be present, and a folder is treated as done only if all expected keys are populated.

The per-folder outputs are seven JSON files (`charge`, `bond`, `qtaim`, `fuzzy_full`, `other`, `orca`, `timings`); the first five are Multiwfn-derived and the per-method schema is in Sec. 3. `orca.json` is parsed directly from the ORCA `.out` via a single-pass enum state machine in `qtaim_gen/source/core/parse_orca.py`, chosen over a general parser like `cclib` (Berquist et al., 2024) because it streams the file once with no runtime dependencies and only extracts the fields we need. `timings.json` is a provenance record (per-step wallclock seconds, no hardware metadata).

Once a tier completes, `json-to-lmdb` converts the JSONs into typed LMDB files (one per descriptor family) sharing a common key derived from the relative path of the job folder. The LMDB representation is portable and self-describing, so it is directly usable by external graph models like Chemprop (Heid et al., 2024) or any custom PyG/DGL pipeline. For users of `qtaim_embed`, the `generator-to-embed` entrypoint composes the typed LMDBs into PyG heterographs through `GeneralConverter`, with configurable bonding (cutoff, fuzzy, IBSI, or QTAIM), charge filter, fuzzy filter, and ORCA-globals filter, exposing atom-, bond-, and global-level targets simultaneously.

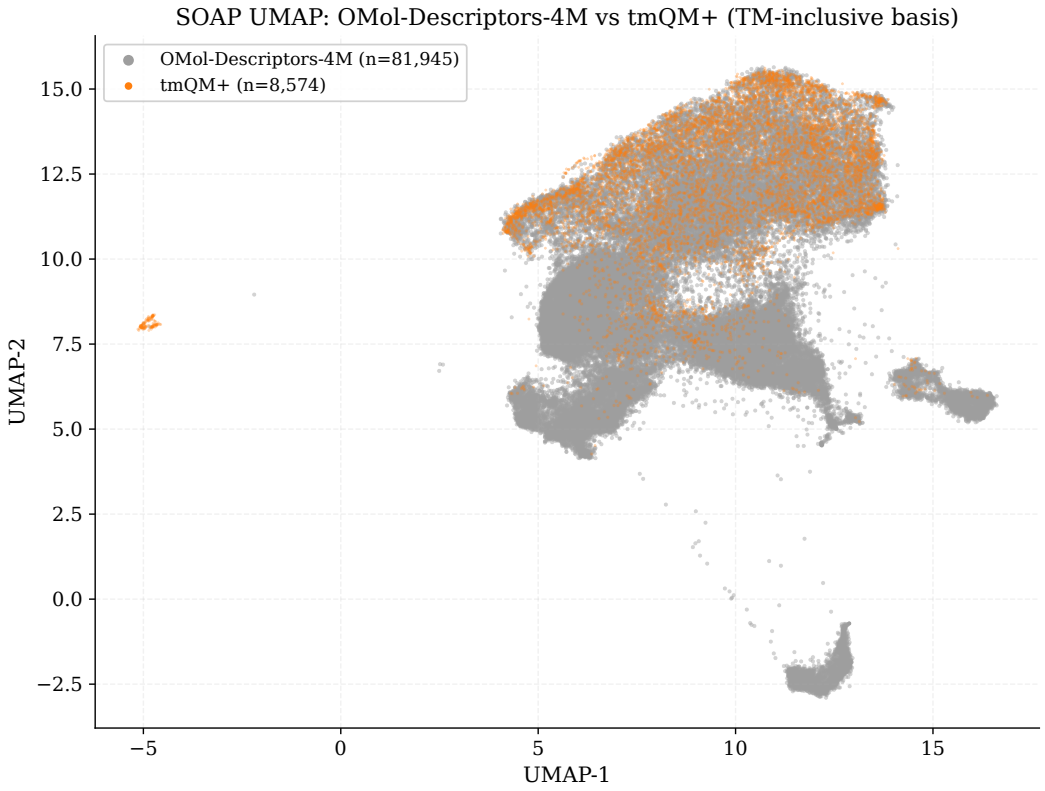

Figure 13: SOAP-UMAP embedding of OMol-Descriptors-4M (gray) against tmQM+ (orange) at the 32-element TM-inclusive SOAP basis. tmQM+ sits largely inside the TM-coordination region (98.3 % of plotted points) of OMol-Descriptors-4M (upper-central cluster, UMAP-1 ≈ 5–14, UMAP-2 ≈ 9–15). The lower gray-only territory (UMAP-2 ≈ 0–7) and the bottom-right tongue (UMAP-2 ≈ -3, UMAP-1 ≈ 12) hold OMol's organic, electrolyte, and protein-pocket chemistry that tmQM+ does not span. The left-hand island (UMAP-1 ≈ -5, UMAP-2 ≈ 8) holds 145 tmQM+ points (1.7 %) and 2 OMol-Descriptors-4M points (`ml_mo`).

O.1 HPC Portability and Reproducibility

We exercised the pipeline across three HPC schedulers (PBS Pro, Slurm, and Flux, at three computing centers whose names are withheld for review). The production run of this dataset used the systems in App. B. PBS and Slurm use a controller-and-workers Parsl (Babuji et al., 2019) topology where a login node farms job folders out to allocated worker nodes via the `PBSProProvider`/`SlurmProvider` executors. Flux uses a single-node Parsl `ThreadPoolExecutor` that works through a local job queue without remote dispatch; this is the same path used for laptop and workstation runs (`full-runner`). All three configs live in `qtaim_gen/source/utils/parsl_configs.py` as minimal edits of the same template, so porting to a new center is expected to require only swapping the provider class and queue/account fields.

The pipeline pins no specific ORCA or Multiwfn version. The reported dataset was produced with ORCA 6.0.0 (matching OMol25) and Multiwfn 3.8; ORCA 5/6.1 and Multiwfn 3.9 have also been exercised. Determinism is structural rather than numeric: the LMDB key derivation, per-tier JSON-key sets, and parsing logic for Multiwfn and ORCA outputs are fixed by code and covered by the test suite under `tests/`. Numeric reproducibility of the underlying QM calculation is the responsibility of ORCA. Running `pytest -q` on a fresh checkout exercises json-to-lmdb, sharded conversion, and merging end-to-end, so a user's local toolchain reproduces the LMDB and JSON artifacts we ship.

P RUNNING A SINGLE JOB FOLDER

The same entry point that drives production runs processes one folder. The folder needs an ORCA wavefunction (`.gbw`) and the ORCA input that produced it; an `orca.out`, if present, is parsed into `orca.json`.

```
full-runner --folder ./my_job \
  --multiwfn_cmd /path/to/Multiwfn_noGUI \
  --orca_2mkl_cmd /path/to/orca_2mkl --n_threads 4
```

This writes `charge.json`, `bond.json`, `qtaim.json`, `fuzzy_full.json`, `other.json` and `timings.json` into the folder. On the 21-atom example shipped with the repository, the tier-0 run takes about 8 s on 4 threads. Replacing `--folder` with `--job_file` (one path per line) runs a list of folders; the Parsl runners (App. O) distribute that list across nodes.

Q THROUGHPUT AND PARALLEL SCALING

Here we report a local scaling experiment that measures the concurrency performance of Multiwfn on our routines. The local sweep is run on 16 copies of a moderately-sized structure drawn from the dataset (from the droplet vertical here). It contains 36 atoms where we ran the three levels of calculations in triplicate with Multiwfn 3.8. We sweep the number of concurrent job-folder workers $\in \{1, 2, 4, 16\}$ at a fixed budget of cores (16) to measure how concurrency impacts wallclock. The configurations are W16T1 (16 workers $\times$ 1 thread/job), W4T4 (4 workers $\times$ 4 threads/job), W2T8 (2 workers $\times$ 8 threads/job), and W1T16 (1 worker $\times$ 16 threads/job). W16T1 has the lowest median wallclock at every tier, but the margin shrinks as the tier grows: W2T8 takes 1.63$\times$ as long as W16T1 at tier 0 and 1.06$\times$ at tier 2, where W16T1, W4T4 and W2T8 fall within 6 % of each other (3,416, 3,539 and 3,635 s). Only W1T16 is several times slower at every tier. We therefore recommend many single-threaded workers for tiers 0 and 1. At tier 2, and for large systems where 16 concurrent Multiwfn processes may exceed node memory (memory was not measured here), W4T4 or W2T8 cost little throughput.

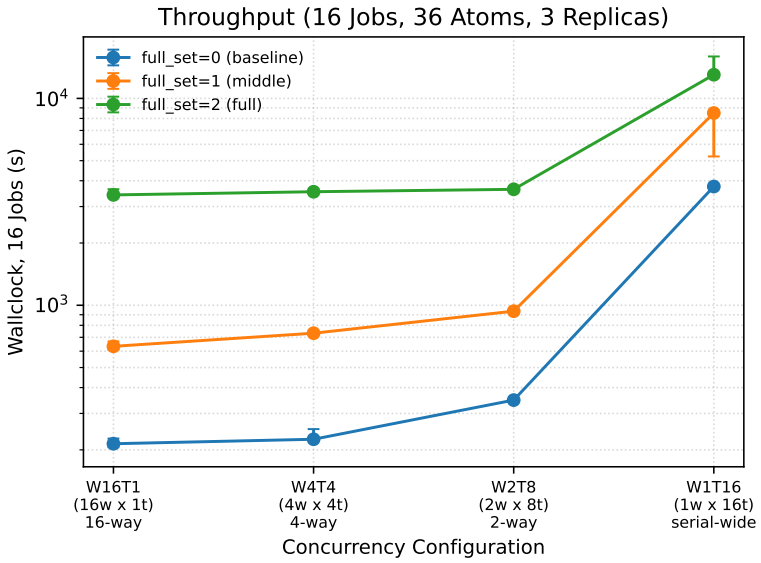

Figure 14: Multiwfn QTAIM post-processing wallclock as a function of intra-node concurrency configuration. Each configuration processes 16 homogeneous copies of `droplet_test` (36 atoms, def2-TZVPD) under cascade-restart tiers. Configurations saturate a 16-thread budget differently: W16T1 (16 workers × 1 thread), W4T4 (4 workers × 4 threads), W2T8 (2 workers × 8 threads), W1T16 (1 worker × 16 threads). Each curve represents a calculation tier (`full_set` 0=baseline, 1=middle, 2=full). Tiers 1 and 2 use `--restart` from the prior tier, so per-tier totals are cumulative phase wallclocks. Markers show medians across 3 replicates with min/max whisker error bars. Y-axis is logarithmic to visualize the 75× range spanning W16T1 tier 0 (fastest) to W1T16 tier 2 (slowest). Hardware: AMD 7950X (16c/32t). ORCA 6.0.1 (`shared_openmpi416`) with system OpenMPI 4.1.2; Multiwfn 3.8 (Oct 2024 dev). ORCA SCF is precomputed; this measures post-processing only.