DEMIR: DENSITY-MATRIX-INFORMED REPRESENTATIONS FOR MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The Hohenberg–Kohn theorem, which states that all physical quantities in the ground state are functionals of the electron density, suggests that much of a molecule's behavior is encoded in its electronic structure. This study proposes DeMiR, a model that treats electron density matrix prediction as a pretraining task. DeMiR reconstructs cc-pVDZ/B3LYP density matrices for molecules ranging in size from QM9-scale molecules to drug-like molecules with up to 146 atoms, achieving median NMAE values of 1.0% and 2.16%, respectively. At the atomic level, the model's embeddings contain electronic structure information: per-atom electronic observables are predicted with an R^2 of 0.94 by a linear probe, whereas Uni-Mol reaches 0.84 and a randomly initialized model of the same architecture only 0.63. The embeddings also remove 44.9% of the residual error left by a geometry-based descriptor, showing that this information is not linearly accessible from an explicit geometric descriptor. At the molecular level, transfer depends on the label type. Where electronic structure directly determines the property, transfer holds: on an external experimental pKa test set, fine-tuned DeMiR achieves an MAE of 1.143, significantly outperforming a randomly initialized control with the same architecture and training budget (1.397), and performing on par with or better than chemical foundation models pretrained on far larger datasets. In contrast, for labels such as pIC50 and ADMET, DeMiR shows no consistent advantage over existing general-purpose foundation models, as two-dimensional substructures and physicochemical descriptors have already saturated predictive performance. This study presents the scope within which a connection between electron density prediction capability and molecular property prediction capability holds, along with its boundaries.

1 INTRODUCTION

Since the early days of cheminformatics, when molecules first began to be represented abstractly on computers, how to represent molecules has remained a major problem. Attempts to create embeddings that effectively represent molecules have continuously evolved. SMILES represents molecules as strings (Weininger, 1988), ECFP constructs molecular substructures based on fixed rules and vectorizes them (Rogers & Hahn, 2010), other approaches represent molecules through their physicochemical properties (Todeschini & Consonni, 2009), and chemical foundation models are trained with their own pretraining tasks, such as ChemBERTa (Chithrananda et al., 2020) and Uni-Mol (Zhou et al., 2023). More recently, the internal representation of a co-folding model such as Boltz-2 (Passaro et al., 2025) has also been investigated for small-molecule learning (Jang et al., 2026).

According to the Hohenberg–Kohn theorem, all physical properties in the ground state are uniquely determined by the functional of the ground-state electron density (Hohenberg & Kohn, 1964). Meanwhile, under the Born–Oppenheimer approximation (Born & Oppenheimer, 1927), the ground-state electron density is also determined by the atom type, its 3D coordinates, charge, and spin state. Therefore, in principle, this information is already contained in the 3D structural representation. However, the theorem only guarantees the existence of such a functional. It says nothing about whether that information is extractable or whether it can be learned from only a limited set of labels.

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.

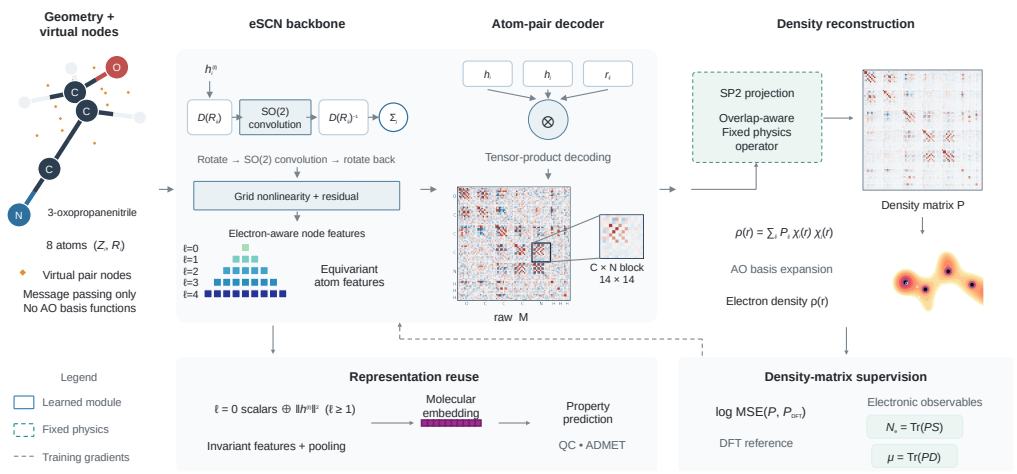


Figure 1: Overview of DeMiR. An equivariant eSCN encoder maps atomic identities and three-dimensional coordinates to atom-wise features. During pretraining, an atom-pair decoder and an overlap-aware SP2 projector reconstruct the AO density matrix, from which the electron density is evaluated. Encoder features are reused for downstream prediction without evaluating the density-matrix head. The invariant feature and pooling pathway is illustrated. Dashed arrows indicate training gradients.

In this regard, several studies have been conducted to use electronic structure as input for representing molecules (Qiao et al., 2020; Boiko et al., 2025; Suriana et al., 2025; Zhang et al., 2026). However, since calculating quantum chemical properties—including electronic structure—using DFT is computationally expensive, research focused on predicting these properties has also been pursued as a separate line of inquiry (Kaniselvan et al., 2026; Kim et al., 2024). Closest to our study, dm-PhiSNet (Yescas-Ramos et al., 2026) also predicts atomic-orbital density matrices with an equivariant network followed by a purification step, but uses them to accelerate SCF calculations rather than to represent molecules. To bridge these approaches, Fallani et al. (2025) improved ADMET modeling performance by using quantum property prediction as a pretraining task. However, this study compared performance only with the random-init model of the same architecture. It also limited performance comparisons to the same architecture when different quantum chemical properties were used as pretraining tasks. Since it did not compare the approach to other molecular representations, a controlled comparison is needed to determine what quantum chemical supervision actually adds to the molecular representation. Our contributions are as follows:

- We propose DeMiR (Fig. 1), a single transferable model that predicts idempotent, electron-conserving density matrices for molecules ranging from QM9 to drug-like size, with the physical constraints enforced inside the training loop.
- We show that its atom embeddings encode electronic structure beyond what an explicit geometric descriptor captures, outperforming Uni-Mol, Boltz-2 and a random-init control on per-atom electronic observables.
- This information transfers to molecular properties governed by electronic structure: frozen DeMiR embeddings outperform external representations on 14 of 15 QM9 targets held out from pretraining, and fine-tuned DeMiR outperforms a matched random-init control on experimental pKa.

2 PRELIMINARIES

In Kohn–Sham DFT (Kohn & Sham, 1965), the ground-state electron density of an N_e-electron molecule is obtained from auxiliary single-particle orbitals ψ_i instead of the $3N_e$-dimensional many-electron wave function. Expanding each orbital in atom-centered basis functions χ_μ (LCAO),

$$\psi_i(\mathbf{r}) = \sum_\mu c_{\mu i}\, \chi_\mu(\mathbf{r}), \tag{1}$$

the density is fully specified by the density matrix

$$P_{\mu\nu} = \sum_{i}^{\text{occ}} f_i \, c_{\mu i} \, c^*_{\nu i}, \quad \left(n(\mathbf{r}) = \sum_{\mu\nu} P_{\mu\nu} \, \chi_\mu(\mathbf{r}) \, \chi^*_\nu(\mathbf{r}) \right). \tag{2}$$

In a closed-shell molecule every occupied orbital holds two electrons ($f_i = 2$), so a valid density matrix is idempotent and conserves the electron count,

$$\mathbf{PSP} = 2\mathbf{P}, \qquad \mathrm{Tr}(\mathbf{PS}) = N_e, \tag{3}$$

where $S_{\mu\nu} = \langle \chi_\mu | \chi_\nu \rangle$ is the overlap matrix of the non-orthogonal basis. In the Löwdin-orthogonalized basis, $\mathbf{Q} = \mathbf{S}^{1/2}\mathbf{P}\mathbf{S}^{1/2}$, these conditions become $\mathbf{Q}^2 = 2\mathbf{Q}$ and $\mathrm{Tr}\,\mathbf{Q} = N_e$, and the diagonal elements $Q_{\mu\mu}$ are Löwdin populations. Finally, when the molecule is rotated, basis functions of angular momentum ℓ mix through the Wigner-D matrix $\mathbf{D}^{(\ell)}$, so the block of $\mathbf{P}$ between shells of angular momenta a and b transforms as $\mathbf{P}_{ab} \mapsto \mathbf{D}^{(a)}\mathbf{P}_{ab}\mathbf{D}^{(b)\mathsf{T}}$. A model that predicts $\mathbf{P}$ from atomic coordinates should therefore be rotation-equivariant, and its output must satisfy Eq. (3).

3 METHODS

3.1 DATASET

Because calculating electron density using DFT is computationally expensive, we calculated electron density for the smallest possible molecules. We used it for pretraining to enable the model to capture electron-density information from the molecule's 3D structure. We calculated the electron density matrices for QM9 (Ramakrishnan et al., 2014) molecules with nine or fewer heavy atoms using DFT with PySCF (Sun et al., 2018). Since using QM9 alone limits the types of elements to H, C, N, O, and F, we selected a representative set of molecules from QMugs (Isert et al., 2022) to train the model on additional elements—such as S, P, Cl, and Br—that can be found in drug-like molecules, as well as on larger molecules. Specific details about the dataset can be found in Appendix C.

3.2 MODEL ARCHITECTURE

DeMiR consists of a molecular encoder equivariant to molecular rotation, a decoder that predicts density matrix elements from atom pairs, and an overlap-aware projection module. The encoder maps the types and 3D coordinates of atoms within a molecule to an atom-wise representation. During pretraining, the decoder maps this atom-wise representation to the spin-summed Kohn–Sham density matrix $\widehat{\mathbf{P}}$ in the atom-centered cc-pVDZ atomic-orbital (AO) basis, from which the real-space density follows by Eq. (2).

Molecular graph and virtual nodes. We augment the atomistic graph by placing auxiliary nodes at the midpoints of atom pairs in the molecular graph. Inspired by the use of off-atom representations in electron density prediction (Fu et al., 2024), we introduce one virtual node for each unordered pair of real atoms (i, j) satisfying $0 < \|\mathbf{r}_i - \mathbf{r}_j\|_2 \leq 3\,\text{Å}$, placed at the midpoint $(\mathbf{r}_i + \mathbf{r}_j)/2$. Each virtual node has a learnable type embedding distinguished by the atomic numbers $\{Z_i, Z_j\}$ of the two atoms that compose it. These distance-based virtual nodes are not limited to covalent bonds. They also cover cases where atoms are not bonded but are close enough to influence each other's electronic environment. Virtual nodes take part only in the message passing process and the density matrix decoder operates only on real node representations.

Equivariant spherical-channel encoder. We use the efficient spherical-channel network (eSCN) (Passaro & Zitnick, 2023) as the encoder: each node carries real spherical-harmonic features $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^{(2\ell+1)\times C}$ for degrees $\ell = 0, \ldots, 4$ with $C = 128$, which rotate with the corresponding Wigner-D matrices, so the representation is $\mathrm{SO}(3)$-equivariant. The encoder stacks four message-passing layers of the standard eSCN form: node features are rotated into each edge's frame, convolved with an $\mathrm{SO}(2)$ kernel truncated at $|m| \leq 2$, and rotated back for aggregation, with a Gaussian expansion of the node distance modulating the message and a pointwise SiLU applied on a spherical grid (Passaro & Zitnick, 2023).

Atom-pair density-matrix decoder. The decoder first applies an equivariant linear projection that reduces the encoder features to 16 channels per degree. For each real-atom pair, learned tensor products combine the projected representations into features associated with AO shell pairs. For shells with angular momenta a and b, coupled features are constructed for all degrees $J = |a-b|, \ldots, a+b$. The cc-pVDZ basis contains s, p and d shells, so $J \leq 4$, and a separate module serves each of the six angular shell-pair types.

For radial shell indices s and t, an AO block is reconstructed as

$$A_{i,s,a,m_a;\,j,t,b,m_b} = \sum_{J=|a-b|}^{a+b} \sum_{M=-J}^{J} C^{JM}_{am_a,bm_b} w_{ab,J}(d_{ij})\, u^{JM}_{ij,st}, \tag{4}$$

where $d_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2$, $u^{JM}_{ij,st}$ is a learned tensor-product output, and $C^{JM}_{am_a,bm_b}$ denotes the Clebsch–Gordan transformation in the real spherical-harmonic convention used by the implementation. For each angular shell-pair type, a distance-dependent multilayer perceptron with one 64-unit hidden layer and a SiLU activation produces the scalar weights $w_{ab,J}$, shared across magnetic components and radial shell pairs.

We predict intra-atomic blocks at $d_{ii} = 0$ and interatomic blocks for real-atom pairs within 5 Å. The blocks are assembled in the PySCF AO ordering and symmetrized, $\mathbf{M} = (\mathbf{A} + \mathbf{A}^{\mathsf{T}})/2$, with padded basis entries masked out. By construction the raw matrix transforms covariantly under rotation, $\mathbf{M} \mapsto \mathbf{D}_{\mathrm{AO}} \mathbf{M} \mathbf{D}_{\mathrm{AO}}^{\mathsf{T}}$, where $\mathbf{D}_{\mathrm{AO}}$ is the block-diagonal rotation representation of the AO basis.

Density-matrix projection. The symmetrized raw matrix $\mathbf{M}$ is not yet a valid density matrix. We map it to one with trace-correcting second-order (SP2) purification (Niklasson, 2002) in the Löwdin-orthogonalized metric of Section 2: starting from a spectrally rescaled $\widetilde{\mathbf{S}}^{1/2}\mathbf{M}\widetilde{\mathbf{S}}^{1/2}$, where $\widetilde{\mathbf{S}}$ is a regularized overlap matrix, the recursion drives the matrix to idempotency while steering its trace to $N_e/2$, so that the final $\widehat{\mathbf{P}}$ satisfies the closed-shell conditions of Eq. (3) up to small numerical residuals (Appendix A). The recursion is stopped once the idempotency residual falls below 10^{-5}. This projection uses only matrix multiplication and does not involve eigenvalue decomposition of the predicted matrix. The network therefore learns only the occupied subspace: the electron count is fixed by the projection, and the auxiliary electron-count loss (Appendix B) only penalizes residual deviations. Details of the density-matrix projection are given in Appendix A.

Pretraining. The model is trained on the projected density matrix with a matrix-element loss and auxiliary real-space density, dipole and electron-count terms, and the checkpoint with the lowest unweighted validation matrix error is used throughout (Appendix B).

Representations for downstream prediction. We extract the final real-atom encoder features before the density-matrix decoder. Each atom is represented by a 640-dimensional rotation-invariant vector, the 128 scalar channels $\mathbf{h}_i^{(0)}$ concatenated with the per-channel squared norms $\sum_m (h^{(\ell)}_{i,m,c})^2$ of degrees $\ell = 1, \ldots, 4$. Atom-level probes use these vectors directly, and for molecule-level tasks they are aggregated by mean or sum pooling or by learned attention-based pooling (Ilse et al., 2018). Molecule-level features are extracted without virtual nodes, as are the all-atom geometry controls of Table 1, whereas the other atom-level probes use the pretraining graph. Feature extraction requires only atomic identities and three-dimensional coordinates, without a density-matrix prediction or a self-consistent-field calculation.

4 Results

We evaluate DeMiR at three levels: how accurately it reconstructs the electron density (Section 4.1), what its atom embeddings encode (Section 4.2), and when that information transfers to molecular properties (Section 4.3).

4.1 Density reconstruction

On 400 QM9 and 200 QMugs held-out molecules, DeMiR reconstructs the real-space density with a median normalized mean absolute error (NMAE) of 1.00% on QM9 and 2.16% on QMugs (Table 4).

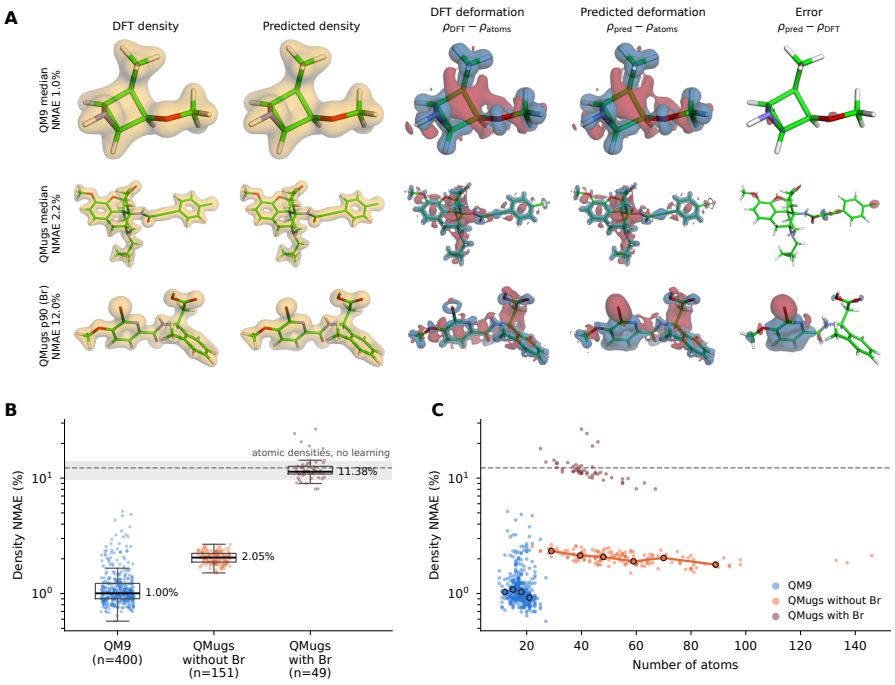


Figure 2: Density reconstruction. (A) DFT and predicted total density (isovalue 0.05 e/a_0^3), deformation density $\rho - \rho_{\text{atoms}}$ and error density (± 0.01 e/a_0^3, blue accumulation and red depletion) for the test molecules at the QM9 median, QMugs median and QMugs 90th percentile of NMAE. (B) Per-molecule NMAE. (C) NMAE against size with binned medians. Dashed line and band: atomic superposition, median and 5–95% range.

A superposition of spherical atomic densities, which contains no bonding information, already scores 12.7% and 11.2%, so the chemically informative deformation density occupies only the last tenth of this scale. DeMiR recovers most of it, without degradation with molecular size up to 146 atoms (Appendix D). An exception to this trend was observed in molecules containing Br—the only period-4 element included in QMugs among those covered by cc-pVDZ basis sets—where the NMAE exceeded 8% for all molecules in the test set containing Br.

4.2 Atom-level electronic information

We probe the frozen atom embeddings with ridge regression for six electronic observables computed from the reference density—Löwdin charge, Mayer valence, maximum Mayer bond order (Mayer, 1983), p- and d-shell population fractions and the magnitude of the on-atom polarization—on 2,000 QM9 molecules, and report the macro-averaged R^2 (Appendix E).

DeMiR embeddings decode these observables with a macro R^2 of 0.939, against 0.842 for Uni-Mol, 0.800 for Boltz-2 and 0.634 for the same architecture at random initialization, and the ordering holds for every target (Table 1, Fig. 3A). Because a randomly initialized equivariant encoder is already a nonlinear function of the geometry, the gap to it can be attributed to density-matrix pretraining. Pretraining exposure has no detectable effect on the result: the probe scores molecules unseen by the encoder as well as seen ones (Appendix E).

This information goes beyond what an explicit geometric descriptor provides. Added to a geometric descriptor that alone reaches R^2 0.916, DeMiR removes 44.9% of the descriptor's residual error, against 17.4% for Uni-Mol, 16.8% for Boltz-2 and 14.7% for random initialization (Fig. 3C). Among the 16,516 hydrogens bonded to carbon, which share element and first-shell composition, the top two principal components of the DeMiR embedding alone recover the on-atom polarization (R^2 0.67, against 0.26 for Uni-Mol and -0.06 for random initialization, Fig. 3B): electronic state lies along the dominant axes of the representation.

Table 1: Atom-level probes, macro R^2. Heavy atoms: all four models compared, encoder run with virtual nodes. All atoms: geometry controls, encoder run without virtual nodes (Boltz-2 has no hydrogen tokens). The two blocks score different atom sets and are not comparable.

Scored atoms	Input geometry	DeMiR	Uni-Mol	Boltz-2	Random init
Heavy	DFT	**0.939**	0.842	0.800	0.634
All	DFT	**0.965**	0.842	—	0.470
All	RDKit	**0.935**	0.824	—	0.457

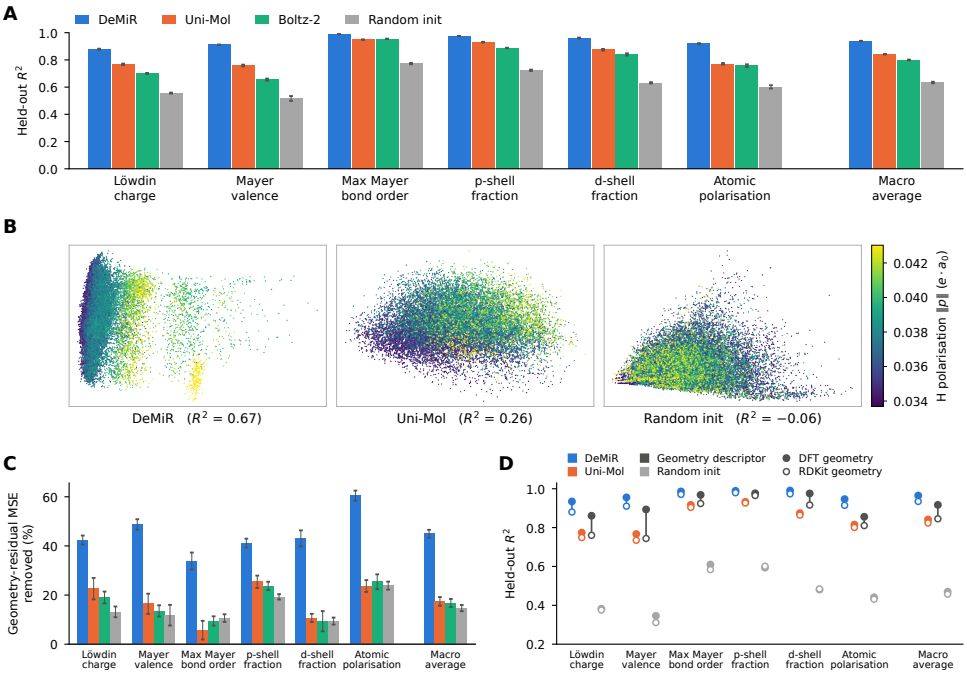


Figure 3: Atom-level evaluation. (A) Probe R^2 per target, four models on the same heavy atoms, with error bars showing SD over five repeats. (B) C–H hydrogens projected onto the top two principal components of each representation, colored by on-atom polarization, with R^2 from the two coordinates alone. (C) Share of a geometric descriptor's residual error removed by each representation. (D) DFT (filled) versus RDKit (open) input geometry.

The information is also accessible under practical constraints. With RDKit conformers instead of DFT geometries, DeMiR loses 0.03 in R^2 while the geometric descriptor loses 0.07, and DeMiR on RDKit geometry still exceeds the descriptor on DFT geometry (Fig. 3D). The advantage is largest with few labels: trained on 25 labeled molecules, a probe on DeMiR embeddings recovers carbon polarization better than a probe on Uni-Mol embeddings trained on 1,000 (Appendix E, Fig. 5).

4.3 Molecule-level transfer

At the molecule level, transfer depends on the label: it holds where electronic structure governs the property and shows no consistent advantage over strong baselines on ADMET tasks. Unless stated otherwise, the encoder is frozen, the pooling and prediction head (k-NN, linear or MLP) are selected on validation data only, and each test set is scored once (Appendix F).

QM9 properties. We probed frozen embeddings on the 15 standard QM9 targets and the four atomization energies, using the QM9 part of the pretraining split so that the 13,153 test molecules were never seen by the encoder. No QM9 property label was used in pretraining, and DeMiR and Uni-Mol received the same RDKit (RDKit contributors) conformers (Riniker & Landrum, 2015). DeMiR outperforms every external representation on 14 of the 15 standard targets (Table 2). On the

Table 2: QM9 frozen probes, test MAE on the 13,153 molecules held out from pretraining (lower is better, best per row in bold). Pooling and head (k-NN, linear, MLP) are selected on validation per representation and target, and MLP results average three seeds. The last four rows are atomization energies. Rand. init: same architecture, random weights. Comp.: element-count vector. †Rotational constant A is dominated by a few outlier molecules in the training and validation sets.

Target	Unit	DeMiR	Rand. init	Comp.	Uni-Mol	MoLFormer	ChemBERTa	Morgan
HOMO	meV	**116**	159	397	187	177	188	162
LUMO	meV	**127**	199	745	237	201	234	182
Gap	meV	**169**	252	797	296	254	283	228
μ	D	**0.575**	0.643	0.973	0.682	0.680	0.712	0.623
α	a_0^3	**0.479**	0.696	2.02	1.13	1.68	1.76	2.40
$\langle R^2 \rangle$	a_0^2	**33.7**	43.6	155	56.8	63.1	61.0	74.0
ZPVE	meV	**7.97**	12.0	37.1	82.5	123	141	170
C_v	cal/(mol K)	**0.192**	0.377	1.62	0.670	0.711	0.745	1.07
A†	GHz	0.595	5.27	0.901	0.632	**0.450**	9.92	40.7
B	GHz	**0.121**	0.160	0.299	0.150	0.153	0.166	0.215
C	GHz	**0.0711**	0.0887	0.200	0.0833	0.103	0.118	0.147
U_0	Ha	0.185	0.0334	**0.0312**	3.17	6.87	7.63	11.1
U	Ha	0.185	0.0334	**0.0311**	3.04	6.86	7.72	11.2
H	Ha	0.185	0.0334	**0.0311**	3.12	6.83	7.61	11.1
G	Ha	0.185	0.0334	**0.0313**	3.00	6.92	7.82	11.1
U_0^{atom}	kcal mol^{-1}	**3.39**	6.32	18.6	22.3	37.1	41.5	56.8
U^{atom}	kcal mol^{-1}	**3.47**	6.26	18.6	22.6	37.4	41.8	57.3
H^{atom}	kcal mol^{-1}	**3.47**	6.19	18.6	22.7	38.0	42.3	57.6
G^{atom}	kcal mol^{-1}	**3.15**	6.11	18.7	20.5	34.6	38.4	53.1

frontier-orbital energies its error is 0.70–0.74 of the best external one, and 0.59–0.63 with attention pooling. The exception is rotational constant A, for which two linear molecules in QM9 have values more than 10,000 times larger than those of the other molecules. The total energies, which are almost linear in the number of atoms of each element, are predicted best by a composition vector and the randomly initialized encoder. Once this atomic contribution is removed, DeMiR is best on all four atomization energies (3.4 kcal/mol for U_0, against 6.3 for random initialization and 22.3 for Uni-Mol). With the same attention pooling, DeMiR also outperforms the identically trained random initialization on all 19 targets (Appendix G).

Physical properties. On four experimental properties governed by the charge distribution—pKa (Baltruschat & Czodrowski, 2020), hydration free energy (Mobley & Guthrie, 2014), gas-to-hexadecane partitioning (Hille et al., 2019) and the thermal decomposition onset temperature of MOFs (Nandy et al., 2022)—frozen DeMiR embeddings outperform Uni-Mol, MoLFormer (Ross et al., 2022) and ChemBERTa on all four (Table 5, Fig. 4A). Uni-Mol receives the same conformers, so the margin is not an effect of three-dimensional input as such.

Transfer on pKa. For pKa, the only one of these tasks with an external test set (278 molecules), we tested whether the advantage is transfer from pretraining. Fine-tuned with learned pooling, DeMiR reaches an MAE of 1.143 ± 0.026, against 1.397 ± 0.024 for the identically trained, randomly initialized control (paired difference -0.254, 95% CI $[-0.361, -0.146]$). It outperforms CheMeleon (Burns & Green, 2025) fine-tuned natively (1.291) and matches Uni-Mol (1.157) (Table 3, Fig. 4C). The two models are close on validation and separate only on the external set, so the gain is consistent with transfer rather than fit, and even the frozen DeMiR encoder with learned pooling (1.221) beats the fully fine-tuned randomly initialized network.

How much of this information reaches the molecular representation depends on the pooling. Under mean pooling, element composition linearly explains 78% of the variance of the pooled vector, and a randomly initialized encoder attains 0.96 of the pretrained encoder's improvement over a trivial baseline (median over fifteen tasks). With learned pooling, this ratio falls to 0.43 on pKa, and the error of the frozen DeMiR encoder decreases by 24% (Appendix F).

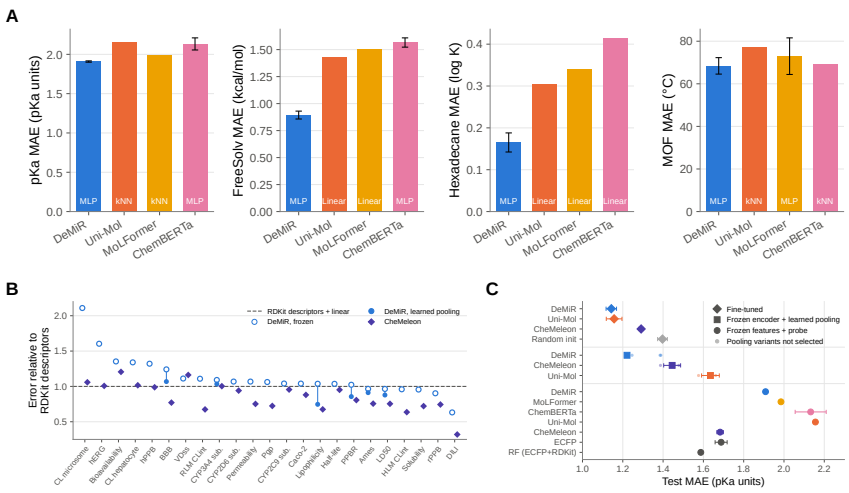


Figure 4: Molecule-level tasks. (A) Physical-property tasks with frozen embeddings. Bar labels give the selected head, and error bars show SD over three seeds of the MLP head. (B) The 23 ADMET tasks of the CheMeleon benchmark, error relative to a linear probe on RDKit descriptors, each task scored with its benchmark metric (MAE or MSE, or one minus AUROC, PRAUC or correlation). Values below 1 beat the descriptor probe. (C) pKa, all models and protocols. Error bars show SD over three seeds, and faint dots show pooling variants not selected on validation.

Table 3: Experimental pKa, external test set ($n = 278$), MAE. The frozen-probe rows of the four representation models are the pKa column of Table 5. $\pm$: SD over three seeds. Rows without it use a deterministic k-NN head.

Protocol	Model	Test MAE
Fine-tuned	DeMiR + learned pooling	**1.143** $\pm$ 0.026
	Uni-Mol (native)	1.157 $\pm$ 0.039
	CheMeleon (native)	1.291 $\pm$ 0.012
	Random init + learned pooling	1.397 $\pm$ 0.024
Frozen encoder	DeMiR + learned pooling	1.221 $\pm$ 0.003
	CheMeleon + learned pooling	1.445 $\pm$ 0.042
	Uni-Mol + learned pooling	1.635 $\pm$ 0.044
Frozen features + probe	Morgan + RDKit descriptors (RF)	1.587 $\pm$ 0.006
	CheMeleon	1.683 $\pm$ 0.016
	ECFP + MLP	1.688 $\pm$ 0.030
	DeMiR, mean pooling	1.909 $\pm$ 0.008
	MoLFormer	1.985
	ChemBERTa	2.132 $\pm$ 0.077
	Uni-Mol	2.156

ADMET. On ADMET tasks, DeMiR shows no consistent advantage over CheMeleon (Burns & Green, 2025), a strong general-purpose molecular foundation model. This is not because the encoder failed to adapt: fine-tuning on the twelve ADMET tasks changed the encoder weights as much as fine-tuning on pKa (median relative L_2 change 1.5×10^{-2} versus 1.7×10^{-2}). Yet fine-tuned DeMiR outperforms CheMeleon on only two of the twelve tasks (half-life and VDss), and the frozen encoder with learned pooling already exceeds it on the same two. Against the identically fine-tuned randomly initialized encoder, pretraining improves two of the twelve tasks significantly (Ames, AUROC $+0.046$, 95% CI $[0.022, 0.071]$, and permeability, Pearson r $+0.091$, $[0.052, 0.132]$) and no significant degradation was detected. On the 23 ADMET tasks of the CheMeleon benchmark, a linear probe on RDKit descriptors beats our frozen representation on 17 and CheMeleon on 6 (Fig. 4B): their predictable signal is largely captured by two-dimensional descriptors. Across the 56-task benchmark of Burns & Green (2025), the frozen representation beats ChemBERTa and Uni-Mol on 48 and 42 tasks but Morgan fingerprints on only 15 and CheMeleon on none, with the losses concentrated in

activity-cliff tasks (Appendix F). Specifically, DeMiR beats them on 1 of the 28 MoleculeACE (van Tilborg et al., 2022) tasks but on 14 of the remaining 28.

5 Conclusion

In this study, we propose DeMiR, a molecular representation model that uses electron density matrix prediction as a pretraining task. DeMiR's atomic embeddings were able to capture electronic observables—beyond what explicit geometric descriptors provide—more accurately than existing molecular representations, and this information was transferred to molecular properties determined by electronic structure, such as pKa, through learned pooling. In contrast, we observed no consistent advantage over strong baselines for ADMET properties, where prediction performance had already plateaued with two-dimensional descriptors. This suggests limitations in explaining these properties—which result from the overlap of multiple biological processes—using the electronic representation of a single molecule alone. However, generating DFT labels is costly, which limited the scale and chemical coverage of the pretraining data, and reconstruction errors were particularly large for bromine-containing molecules. Furthermore, because the method assumes neutral closed-shell molecules and a fixed basis set, additional validation is required to generalize it to broader systems. Given that it is possible to reconstruct electronic density at the desired resolution from the density matrix and use it as a representation, we hope this study will serve as a bridge connecting quantum chemical information with research on drug-like molecules.

AI use statement

Generative AI tools were used in this work for figure-plotting and analysis code, for drafting and editing prose, and for summarizing our own experimental logs. They were not used to generate research ideas, experimental results or data. All AI-assisted code was executed and its outputs were checked against the recorded experimental artifacts by the authors, and every number reported in this paper was traced back to a result file. The authors take responsibility for the final content of this work.

Reproducibility statement

The pretraining data, the density-matrix labels (PySCF, cc-pVDZ/B3LYP) and the train/validation/test splits are described in Section 3.1 and Appendix C. The encoder, decoder and projector are specified in Section 3, with numerical constants in Appendix A, and the probe and fine-tuning protocols in Appendices E and F. All downstream results are obtained with model selection on validation data only. DeMiR denotes the checkpoint with the lowest validation loss of a single pretraining run, and no checkpoint is selected on any downstream result. Code, extracted representations and the analysis scripts that produce every table and figure will be released.

References

Marcel Baltruschat and Paul Czodrowski. Machine learning meets pKa. *F1000Research*, 9:113, 2020. doi: 10.12688/f1000research.22090.2.

Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. doi: 10.1063/1.464913.

Daniil A. Boiko, Thiago Reschützegger, Benjamin Sanchez-Lengeling, Samuel M. Blau, and Gabe Gomes. Advancing molecular machine learning representations with stereoelectronics-infused molecular graphs. *Nature Machine Intelligence*, 7(5):771–781, 2025. doi: 10.1038/s42256-025-01031-9.

M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20): 457–484, 1927. doi: 10.1002/andp.19273892002.

Jackson W. Burns and William H. Green. Deep learning foundation models from classical molecular descriptors. *arXiv preprint arXiv:2506.15792*, 2025.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics*, 90(2):1007–1023, 1989. doi: 10.1063/1.456153.

Alessio Fallani, Ramil Nugmanov, Jose Arjona-Medina, Konstantin Chernichenko, Jörg Kurt Wegner, and Alexandre Tkatchenko. Pretraining graph transformers with atom-in-a-molecule quantum properties for improved ADMET modeling. *Journal of Cheminformatics*, 17:25, 2025. doi: 10.1186/s13321-025-00970-0.

Xiang Fu, Andrew Rosen, Kyle Bystrom, Rui Wang, Albert Musaelian, Boris Kozinsky, Tess Smidt, and Tommi Jaakkola. A recipe for charge density prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 9727–9752. Curran Associates, Inc., 2024. doi: 10.52202/079017-0310.

Christoph Hille, Stefan Ringe, Martin Deimel, Christian Kunkel, William E. Acree, Karsten Reuter, and Harald Oberhofer. Generalized molecular solvation in non-aqueous solutions by a single parameter implicit solvation scheme. *The Journal of Chemical Physics*, 150(4):041710, 2019. doi: 10.1063/1.5050938.

P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2018.

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

Hyosoon Jang, Hyunjin Seo, Honghui Kim, Seonghyun Park, Taewon Kim, Yunhui Jang, and Sungsoo Ahn. A systematic evaluation of co-folding model representations for small-molecule learning. *arXiv preprint arXiv:2602.13249*, 2026.

Manasa Kaniselvan, Benjamin Kurt Miller, Meng Gao, Juno Nam, and Daniel S. Levine. Learning from the electronic structure of molecules across the periodic table. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=PS1YS8Wv4t`.

Jungwoo Kim, Woojae Chang, Hyunjun Ji, and InSuk Joung. Quantum-informed molecular representation learning enhancing ADMET property prediction. *Journal of Chemical Information and Modeling*, 64(13):5028–5040, 2024. doi: 10.1021/acs.jcim.4c00772.

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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.

David L. Mobley and J. Peter Guthrie. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of Computer-Aided Molecular Design*, 28(7):711–720, 2014. doi: 10.1007/s10822-014-9747-x.

Aditya Nandy, Gianmarco Terrones, Naveen Arunachalam, Chenru Duan, David W. Kastner, and Heather J. Kulik. MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Scientific Data*, 9:74, 2022. doi: 10.1038/s41597-022-01181-0.

Anders M. N. Niklasson. Expansion algorithm for the density matrix. *Physical Review B*, 66(15): 155115, October 2002. doi: 10.1103/PhysRevB.66.155115.

Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27420–27438. PMLR, 2023. URL https://proceedings.mlr.press/v202/passaro23a.html.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Zhuoran Qiao, Matthew Welborn, Anima Anandkumar, Frederick R. Manby, and Thomas F. Miller, III. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics*, 153(12):124111, 2020. doi: 10.1063/5.0021955.

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.

RDKit contributors. RDKit: Open-source cheminformatics. https://www.rdkit.org. Version 2026.03.1.

Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55(12): 2562–2574, 2015.

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

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022.

Qiming Sun, Timothy C. Berkelbach, Nick S. Blunt, George H. Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D. McClain, Elvira R. Sayfutyarova, Sandeep Sharma, Sebastian Wouters, and Garnet Kin-Lic Chan. PySCF: the python-based simulations of chemistry framework. *WIREs Computational Molecular Science*, 8(1):e1340, 2018.

Patricia Suriana, Joshua A. Rackers, Ewa M. Nowara, Pedro O. Pinheiro, John M. Nicoloudis, and Vishnu Sresht. Beyond atoms: Evaluating electron density representation for 3D molecular learning. *arXiv preprint arXiv:2511.21900*, 2025.

Roberto Todeschini and Viviana Consonni. *Molecular Descriptors for Chemoinformatics*. Wiley-VCH, Weinheim, 2 edition, 2009. doi: 10.1002/9783527628766.

Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23): 5938–5951, 2022. doi: 10.1021/acs.jcim.2c01073.

David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

Zuriel Y. Yescas-Ramos, Andrés Álvarez-García, and Huziel E. Sauceda. Towards accelerated SCF workflows with equivariant density-matrix learning and analytic refinement. *arXiv preprint arXiv:2604.27256*, 2026.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Wei Zhang, Kun Li, Jiameng Chen, Jiajun Yu, Yizhen Zheng, Duanhua Cao, and Wenbin Hu. UniField: RBF-guided electron density fusion for enhanced molecular representations. *arXiv preprint arXiv:2605.24013*, 2026.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A Density-matrix projection details

Let $S_{\mu\nu} = \langle \chi_\mu \mid \chi_\nu \rangle$ be the AO overlap matrix. Its eigenvalues are floored at 10^{-6} when forming the regularized matrix $\widetilde{\mathbf{S}}$ and its square roots, so the regularized metric equals the original overlap matrix whenever no eigenvalue is clipped. The symmetrized raw prediction $\mathbf{M}$ is moved to the orthonormal representation

$$\widetilde{\mathbf{Q}} = \widetilde{\mathbf{S}}^{1/2}\mathbf{M}\widetilde{\mathbf{S}}^{1/2}, \tag{5}$$

and, with lower and upper Gershgorin spectral bounds L and U, initialized as

$$\mathbf{X}_0 = \frac{\widetilde{\mathbf{Q}} - L\mathbf{I}}{\max(U - L, 10^{-12})}, \tag{6}$$

which preserves spectral order, treating larger eigenvalues of the predicted density proxy as occupied components. With target occupied-orbital count $n_{\mathrm{occ}} = N_e/2$, the recursion is

$$\mathbf{X}_{k+1} = \begin{cases} \mathbf{X}_k^2, & \mathrm{Tr}(\mathbf{X}_k) > n_{\mathrm{occ}}, \\ 2\mathbf{X}_k - \mathbf{X}_k^2, & \text{otherwise}, \end{cases} \tag{7}$$

stopped when $\max_{\mu,\nu} |(\mathbf{X}_k - \mathbf{X}_k^2)_{\mu\nu}| < 10^{-5}$, subject to a cap of 200 updates. The final prediction is

$$\widehat{\mathbf{P}} = 2\widetilde{\mathbf{S}}^{-1/2}\mathbf{X}_\star\widetilde{\mathbf{S}}^{-1/2}, \tag{8}$$

followed by numerical symmetrization and masking. Finite precision and incomplete convergence can introduce residual errors, and overlap regularization can additionally affect the closed-shell identities evaluated with the original $\mathbf{S}$. In particular, the stopping criterion tests idempotency rather than electron-number accuracy. On the held-out density group of Table 4 (400 QM9 and 200 QMugs molecules), every molecule converges before the cap (median 32 and 31 iterations, maximum 36 and 57), no overlap eigenvalue is clipped, and no natural occupation of the output lies between 0.05 and 0.95. Evaluated with the original $\mathbf{S}$, the largest element of $\widehat{\mathbf{P}}\mathbf{S}\widehat{\mathbf{P}}/2 - \widehat{\mathbf{P}}$ is 1.1×10^{-3} (medians 1.2×10^{-4} and 3.2×10^{-4}), and the electron count deviates from N_e by at most 0.06 electrons, a relative error of 2×10^{-4} (medians 0.004 and 0.013 electrons). The overlap matrix is diagonalized once per molecule, but the learned matrix $\widetilde{\mathbf{Q}}$ never is. The decoder cutoff applies to the raw matrix, not to the final density matrix, which can become dense after projection.

B Pretraining details

The loss is $\log \mathrm{MSE}(\widehat{\mathbf{P}}, \mathbf{P}) + w_\rho \log \mathcal{L}_\rho + w_\mu \log \mathcal{L}_\mu + w_N \log \mathcal{L}_N + w_g \mathcal{L}_g$, where $\mathcal{L}_\rho$ is the NMAE of the real-space density on points sampled around the nuclei (500 per atom, at most 50,000), $\mathcal{L}_\mu$ the squared dipole error, $\mathcal{L}_N$ the electron-count error, and $\mathcal{L}_g = (\log\|\widetilde{\mathbf{Q}}\|_F/\sqrt{n})^2$ fixes the output scale, to which the projection is invariant. The weights are ramped linearly: w_ρ from 0.1 to 0.3 over iterations 3,000–8,000, w_μ from 0 to 0.01 and w_N from 0 to 0.02 over iterations 1,500–15,000, with $w_g = 1$. Training uses AdamW (Loshchilov & Hutter, 2019) (learning rate 5×10^{-5}, gradient clipping 5.0, 32 accumulated single-molecule steps) with QM9/QMugs-balanced sampling and a plateau learning-rate scheduler. The encoder and decoder are initialized from a first-stage model of the same architecture trained on the same corpus without the projection. Validation runs every 2,000 iterations, and the checkpoint with the lowest unweighted validation matrix MSE (iteration 6,000) is DeMiR.

C Dataset details

Density-matrix labels are spin-summed Kohn–Sham density matrices of neutral closed-shell molecules computed with PySCF (Sun et al., 2018) at the B3LYP (Becke, 1993)/cc-pVDZ (Dunning,

1989) level on the dataset geometries. The pretraining corpus contains 133,885 QM9 molecules (107,008 training, 13,428 validation and 13,449 test) and 20,781 QMugs molecules (16,724, 2,038 and 2,019), split at the molecule level. The QMugs subset was chosen to add the elements and molecular sizes missing from QM9. From QMugs molecules composed only of H, C, N, O, F, P, S, Cl and Br and containing at least one of P, S, Cl and Br, we selected up to 5,000 molecules per heteroatom by k-means clustering of ChemBERTa SMILES embeddings, taking the molecule closest to each cluster center, and filled the set to 20,000 molecules by the same diversity criterion. A further 781 molecules come from an earlier 1,000-molecule pilot subset selected in the same way.

D DENSITY RECONSTRUCTION DETAILS

NMAE is $\sum_{\mathbf{r}} |\rho_{\mathrm{pred}}(\mathbf{r}) - \rho_{\mathrm{DFT}}(\mathbf{r})| / \sum_{\mathbf{r}} \rho_{\mathrm{DFT}}(\mathbf{r})$ on a uniform grid (VASP-compatible FFT grid, 400 eV cutoff, 4 Å vacuum), evaluated on the first 400 QM9 and 200 QMugs molecules of the test split.

Table 4: Held-out density NMAE (%, lower is better). Atomic superposition: the sum of spherical atomic densities, i.e. the error with no learning.

	n	median	mean	p90
QM9	400	1.00	1.19	1.79
QMugs	200	2.16	4.58	11.67
without bromine	151	2.05	2.06	2.36
with bromine	49	11.38	12.34	15.04
Atomic superposition, QM9	400	12.73	12.68	13.91
Atomic superposition, QMugs	200	11.15	11.03	12.53

The QM9 test molecules have a median of 18 atoms (maximum 27) and the QMugs test molecules a median of 48 (maximum 146). For the molecules closest to the median error of each set, the predicted deformation density $\rho_{\mathrm{pred}} - \rho_{\mathrm{atoms}}$ correlates with the DFT one at 0.996 (QM9) and 0.970 (QMugs) over the grid (Fig. 2A), and no size-dependent degradation is observed (Fig. 2C).

All 49 bromine-containing QMugs test molecules exceed 8% NMAE (8.1–26.6%), while the 151 bromine-free molecules stay below 2.67%. P, S and Cl, the other elements absent from QM9, show no such effect. The error is confined to the bromine atom and the adjacent functional group (Fig. 2A, third row): the predicted density matrix leaves the bromine $3s$, $4s$ and $5s$ shells empty and the $2p$ and $3p$ shells depleted, about 9.8 electrons per atom, and places the charge on the bonded ring while the total electron count is conserved. Forty of the 49 molecules are reconstructed worse than by the atomic superposition. Shells are the radial shells of the cc-pVDZ basis, and populations are Mulliken populations.

E ATOM-LEVEL PROBE DETAILS

Protocol. Probes are standardized ridge regressions ($\alpha = 100$) on the frozen 640-dimensional atom vectors, with splits grouped by molecule (test fraction 0.25) and repeated five times, and R^2 is macro-averaged over targets and elements. All analyses use the same 2,000 QM9 molecules, which every model receives complete, hydrogens included. Only the scored atoms differ. Boltz-2 emits tokens for heavy atoms only, so the four-model comparison scores 17,592 heavy atoms, whereas the geometry controls pair DFT with RDKit coordinates and score all 32,636 atoms of the 1,812 molecules for which RDKit produces a conformer.

Exposure control. The 2,000 molecules were drawn before the final pretraining split was fixed, and 1,583 of them lie in the encoder's training split. Fitting the probe on encoder-training molecules only, DeMiR scores 0.932 on the 206 encoder-test molecules and 0.935 on further encoder-training molecules, a difference no larger than that of Uni-Mol, which never saw these labels (0.833 versus 0.831).

Geometric descriptor. The 162-dimensional descriptor concatenates the composition of each atom's nearest neighbors (10), element-resolved radial distributions (80), sorted nearest-neighbor

distances (8), distance moments (4) and bond-angle moments (60). The residual error removed by a representation is $1 - \mathrm{MSE}(\text{geometry} + \text{embedding})/\mathrm{MSE}(\text{geometry only})$. With RDKit instead of DFT coordinates, the descriptor's R^2 falls from 0.917 to 0.845, DeMiR's from 0.965 to 0.935 and Uni-Mol's from 0.842 to 0.824.

Label efficiency. To remove element identity as a shortcut, this probe scores carbon atoms only, trains on 25 to 1,000 labeled molecules and is evaluated on held-out molecules, with five repeats. DeMiR with 25, 50 and 250 labeled molecules matches or exceeds Uni-Mol with 1,000 on the polarization magnitude (R^2 0.66 versus 0.65), the Mayer valence (0.83 versus 0.78) and the Löwdin charge (0.76 versus 0.74), respectively (Fig. 5).

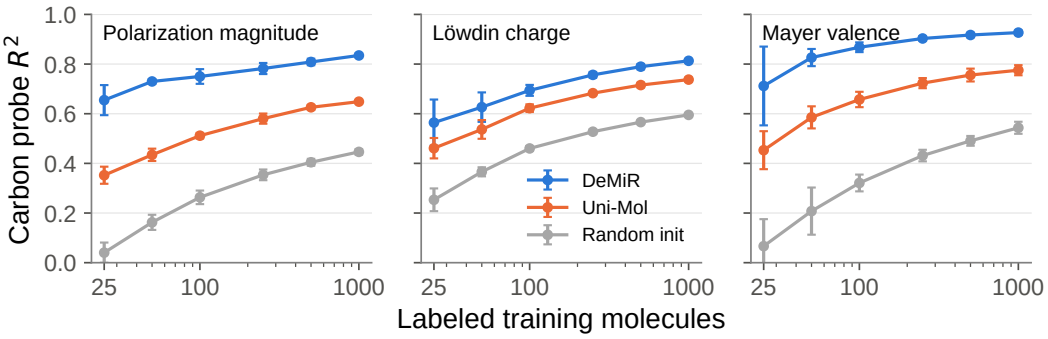


Figure 5: Carbon-atom probe R^2 against the number of labeled training molecules (mean $\pm$ SD over five repeats, ridge on frozen embeddings).

F Downstream details

Table 5: Physical-property tasks, frozen embeddings, test MAE (lower is better). DeMiR and Uni-Mol receive the same RDKit conformers, and MoLFormer and ChemBERTa read SMILES.

	pKa (pKa units)	FreeSolv (kcal/mol)	Hexadecane ($\log_{10} K$)	MOF organic (°C)
DeMiR	**1.909**	**0.894**	**0.165**	**68.4**
Uni-Mol	2.156	1.433	0.304	77.2
MoLFormer	1.985	1.502	0.339	73.0
ChemBERTa	2.132	1.566	0.415	69.2

Table 6 summarizes the evaluation protocols.

Pooling. Unless stated otherwise, learned pooling is gated-attention pooling (Ilse et al., 2018): each 640-dimensional atom vector is projected to 128 dimensions, weighted by a softmax-normalized attention score and summed into a 128-dimensional molecular vector, which a two-layer MLP maps to the target (115k parameters). Where the pooling is selected on validation, DeepSets (Zaheer et al., 2017), PMA (Lee et al., 2019) and an attention variant pooling the raw atom vectors are further candidates. Frozen embeddings are probed with k-NN, linear and MLP heads, and all selection is on validation only.

Fine-tuning. Encoder and pooling are trained end to end with AdamW (separate learning rates, weight decay 10^{-5}, batch 32, gradient clipping 1.0), early-stopped on validation MAE with patience 20 for at most 100 epochs. The randomly initialized control uses the same architecture and protocol. Differences on the pKa test set are paired-bootstrap estimates with 95% confidence intervals.

Pooling comparison. With the encoder frozen, replacing mean pooling by the validation-selected learned pooling improves pKa by 24% (MAE 1.85 to 1.41), whereas the twelve ADMET tasks improve by a median of 1.5%. These runs form a separate pipeline, in which mean pooling is itself a learned layer and the selected pooling is PMA. They are not the frozen rows of Table 3. Element

Table 6: Evaluation protocols. All splits are shared by every representation of a comparison, and all selection uses validation data only. Molecule-level features and the all-atom geometry controls are computed without virtual nodes. The other atom-level probes use the pretraining graph with virtual nodes. [a]Five seeds for tasks with fewer than 500 training molecules.

Evaluation	Data and split	Train / val / test	Encoder and selection	Seeds
Atom-level probes	2,000 QM9 molecules, DFT labels, molecule-grouped splits	75% / – / 25% of molecules	frozen, ridge, $\alpha = 100$ fixed	5 splits
QM9 properties	QM9 part of the pretraining split	104,450 / 13,105 / 13,153	frozen, pooling and head on validation	3 (MLP)
QM9, attention pooling	as above	as above	frozen + attention pooling, learning rate on validation	3
pKa	experimental pKa, parent- and scaffold-grouped validation, fixed external test	4,470 / 1,118 / 278	frozen probe, frozen + learned pooling, or full fine-tuning	3
FreeSolv, hexadecane, MOF	seeded 80/10/10 chemical-parent group splits	503 / 64 / 63, 216 / 27 / 27, 885 / 142 / 61	frozen, pooling and head on validation	3 (MLP)
ADMET fine-tuning	12 CheMeleon benchmark tasks, benchmark splits	per task	full fine-tuning + attention pooling vs. random init, learning rate and budget on validation	3 or 5[a]
ADMET vs. descriptors	23 ADMET tasks of the CheMeleon benchmark, benchmark splits	per task	frozen, pooling and head on validation	3 (MLP)
56-task benchmark	CheMeleon benchmark, benchmark splits	per task	frozen, mean pooling, ridge α on validation	1

composition linearly explains 78% of the variance of the mean-pooled vector, and on the tasks where both were measured, learned pooling lowers this composition R^2 from 0.77 to 0.4–0.5. The ratio of the random initialization's skill over a trivial baseline to that of the pretrained encoder is 0.96 (median over fifteen tasks, 1.03 on pKa) under mean pooling, and 0.94 under learned pooling, with 0.43 on pKa, 0.72 on lipophilicity and 0.91 on the hexadecane partition.

ADMET. Fine-tuning moves the encoder weights by a relative L_2 norm of 1.7×10^{-2} on pKa and by a median of 1.5×10^{-2} (up to 4.6×10^{-2}) on the twelve ADMET tasks. The twelve tasks are those on which an earlier frozen probe came within about 10% of CheMeleon's published value, that is, the tasks most favorable to DeMiR. Fine-tuned DeMiR outperforms CheMeleon on VDss (Spearman 0.529 against 0.401) and half-life (0.444 against 0.327), and the frozen encoder with learned pooling already does so on both (0.517 and 0.417). Two of the twelve tasks have about 100 training molecules. The 56-task comparison uses a wide-range ridge probe on mean-pooled embeddings, and the losses against Morgan fingerprints concentrate in the 28 MoleculeACE activity-cliff tasks. The pretrained and randomly initialized encoders are each fine-tuned once per task. Their confidence intervals are paired bootstraps over test molecules (2,000 resamples) and are not adjusted for multiple comparisons across the twelve tasks.

G QM9 DETAILS

Probe splits are the QM9 part of the pretraining split: after removing molecules without an RDKit conformer or a baseline embedding, 104,450 training, 13,105 validation and 13,153 test molecules remain, identical for every representation. Training and validation molecules were seen during pretraining (geometry and density matrix only), whereas test molecules were not, and no QM9 property was ever used before probing. Atomization energies are the total energies minus the atomic reference energies of QM9. Attention pooling uses the architecture and training protocol of the pKa experiments. For each target, DeMiR and the randomly initialized encoder were trained on the same machine.

Table 7: QM9 with learned gated-attention pooling on frozen atom features, test MAE (mean ± SD over three seeds, learning rate selected on validation). DeMiR and the randomly initialized encoder of the same architecture use the identical pooling and budget and run on the same host for each target. Attention pooling is a weighted mean, so it is ill-suited to size-extensive targets (energies, ZPVE, C_v), for which Table 2 is the relevant comparison.

Target	Unit	DeMiR	Random init	Ratio
HOMO	meV	**101** ± 2	148.0 ± 0.3	0.68
LUMO	meV	**106.7** ± 1.1	175 ± 2	0.61
Gap	meV	**143.6** ± 1.5	228 ± 7	0.63
μ	D	**0.515** ± 0.003	0.6053 ± 0.0014	0.85
α	a_0^3	**0.4719** ± 0.0019	0.68 ± 0.02	0.69
$\langle R^2 \rangle$	a_0^2	**30.1** ± 0.2	38.9 ± 0.8	0.77
ZPVE	meV	**17.7** ± 0.8	22 ± 2	0.80
C_v	cal/(mol K)	**0.212** ± 0.009	0.321 ± 0.008	0.66
A†	GHz	**1.15** ± 0.15	6 ± 3	0.19
B	GHz	**0.1038** ± 0.0018	0.142 ± 0.004	0.73
C	GHz	**0.0590** ± 0.0005	0.091 ± 0.002	0.65
U_0	Ha	**0.98** ± 0.06	1.08 ± 0.08	0.91
U	Ha	**0.98** ± 0.07	1.14 ± 0.09	0.86
H	Ha	**1.04** ± 0.07	1.11 ± 0.03	0.94
G	Ha	**0.99** ± 0.06	1.08 ± 0.04	0.92
U_0^{atom}	kcal mol^{-1}	**6.0** ± 0.5	8.3 ± 0.6	0.73
U^{atom}	kcal mol^{-1}	**6.0** ± 0.4	8.0 ± 0.2	0.75
H^{atom}	kcal mol^{-1}	**6.1** ± 0.3	8.3 ± 0.2	0.73
G^{atom}	kcal mol^{-1}	**5.6** ± 0.2	7.6 ± 0.5	0.74

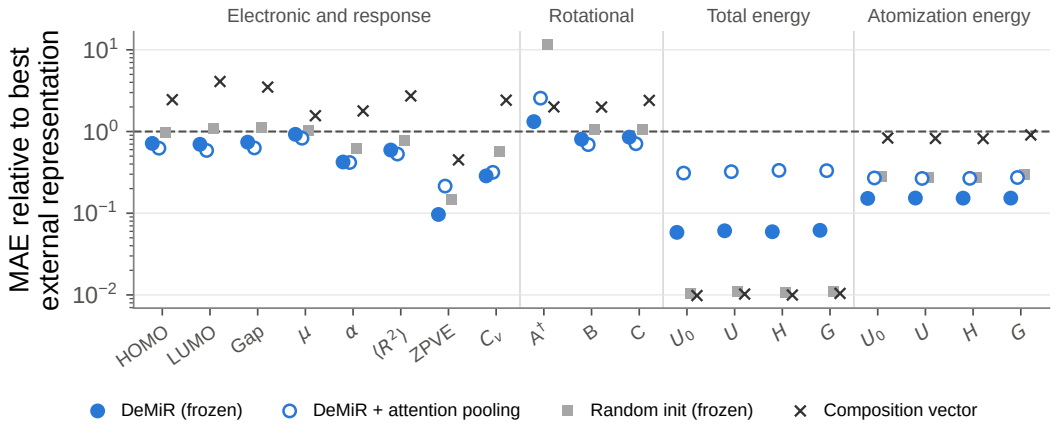


Figure 6: QM9 properties on the 13,153 test molecules held out from pretraining. Test MAE of each representation relative to the best external representation for that target (lowest MAE among Uni-Mol, MoLFormer, ChemBERTa and Morgan fingerprints). Values below the dashed line beat all external representations. Frozen embeddings use a pooling and head selected on validation. Attention pooling is a weighted mean and is therefore disadvantaged on size-extensive targets. Atomization energies are total energies minus atomic reference energies. †Rotational constant A is dominated by a few outlier molecules in the training and validation sets.