Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Accurate quantum chemical property prediction with transferable electron-density representations

Xiangxiang Zeng

xzeng@hnu.edu.cn

Hunan University https://orcid.org/0000-0003-1081-7658

Hongxin Xiang

Hunan University https://orcid.org/0000-0001-8345-8735

Yang Sun

Xiangtan University

Qingrui Liu

Xiangtan University

Shuting Jin

Wuhan University of Science and Technology

Dapeng Xiong

Southeast University https://orcid.org/0000-0002-9337-3839

Jun Xia

The Hong Kong University of Science and Technology (Guangzhou) https://orcid.org/0000-0002-7993-0803

Yanyan Li

Tsinghua University

Jingyu Sheng

Xiangtan University

Tianhang Zhang

Xiangtan University

Tong Wang

Tsinghua University https://orcid.org/0000-0002-9483-0050

Xuan Lin

Xiangtan University https://orcid.org/0000-0001-7216-1821

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11439055/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Accurate quantum chemical property prediction with transferable electron-density representations

Hongxin Xiang[1†], Yang Sun[2†], Qingrui Liu[2], Shuting Jin[3], Dapeng Xiong[4], Jun Xia[5,6], Yanyan Li[7], Jingyu Sheng[2], Tianhang Zhang[2], Tong Wang[8✉], Xuan Lin[2✉], and Xiangxiang Zeng[1✉]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, China
[2]College of Computer Science, Xiangtan University, Xiangtan, 411105, China
[3]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, China
[4]State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing, 211189, China
[5]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China
[6]The Hong Kong University of Science and Technology, Hong Kong, 999077, China
[7]Department of Physics, Tsinghua University, Beijing, 100084, China
[8]School of Life Sciences, Tsinghua University, Beijing, 100084, China
✉**Corresponding Authors:** tongwang@mail.tsinghua.edu.cn, jack_lin@xtu.edu.cn, xzeng@hnu.edu.cn
[†]These authors contributed equally to this work

ABSTRACT

Machine-learning models of molecules usually encode nuclear geometry, while spatial electronic information is either omitted or requires costly molecule-specific quantum calculations. Here we present eMol, an electron-density-driven model that overcomes this trade-off by coupling atomic scaffolds with density-derived probes via a retrievable memory, allowing compatible electronic environments to be reused across molecules without repeated first-principles density calculations. To integrate the transferred electronic information with nuclear geometry, eMol employs an equivariant atom–electron architecture that captures atomic geometry, local electronic structure and higher-order many-body interactions within a unified representation. To avoid computing electron density for every query, eMol uses structure-only retrieval and geometric alignment to reuse probes from a memory of approximately two million previously characterized molecules. This design yields consistently strong performance across 24 quantum-chemical property prediction tasks and 24 molecular potential-learning systems, and reproduces reference structural distributions in biomolecular dynamics. Its learned attention patterns align with local electron density, and it prioritizes literature-supported compounds in drug-activity datasets. More broadly, decoupling expensive physical information from real-time computation provides a generalizable paradigm for molecular learning.

Introduction

Accurate quantum chemical prediction is a fundamental task in molecular modelling[1–3], enabling the estimation of molecular properties[4–6], potential energy surfaces[7,8] and atomic forces[9,10] for molecular simulation and drug discovery[11–14]. Traditional quantum chemical calculations provide physically grounded results but are often too costly for large chemical spaces or long-time molecular dynamics[15–18]. Machine learning approaches offer a promising route to accelerate this process by learning predictive models from quantum chemical data[19–22]. Recent molecular models[10,22,23] and equivariant geometric networks[19,24,25] commonly encode molecules as graphs or three-dimensional atomic structures, using atom types, bonds and nuclear coordinates as the primary description of molecular systems[3,16]. Despite their promising performance, these models remain centred on the nuclear scaffold, thus limiting their representation capabilities to molecular descriptions that do not explicitly incorporate electron density. This motivates molecular representations in which local electronic environments can participate directly in spatial interactions, rather than being inferred solely from nuclear geometry.

However, a physically grounded molecular representation should account not only for nuclear geometry, but also for the surrounding electronic distribution[26,27]. Electron density, $\rho(\mathbf{r})$, provides a real-space description of this distribution and is a central quantity in electronic-structure theory[28]. According to the Hohenberg-Kohn theorem, the ground-state electron

density uniquely determines all ground-state molecular properties[29]. These principles suggest that electron density provides a physically grounded complement to atom-centred molecular geometry for learning molecular representations. However, electron density is typically obtained from quantum chemical calculations such as density functional theory (DFT)[30–32], which remain computationally expensive and time-consuming at scale[33]. Consequently, for new or unseen molecules, molecule-specific electron-density information is often unavailable without additional quantum chemical calculation. Recent progress in large-scale electron-density databases[34] opens a different possibility: existing density fields may be reused as transferable local electronic environments, provided that they can be retrieved, aligned and coupled to new molecular structures in a physically meaningful way. This possibility raises two related challenges: (i) how to encode spatial electronic environments jointly with nuclear geometry, and (ii) how to reuse them for new queries without first calculating their molecule-specific densities. A retrieval mechanism based on molecular structure rather than task-specific labels could further make the same electronic-environment acquisition strategy applicable across different prediction objectives.

In this study, we present eMol, an electron-density-driven molecular model for quantum chemical prediction. eMol is built on two central ideas. First, it represents molecules by coupling atomic scaffolds with localized density-derived electron probes in a shared three-dimensional space, allowing molecular representations to include both nuclear geometry and local electronic environments. Each probe carries a spatial position and local density descriptor, enabling electronic information to participate in geometric message passing rather than being appended only as an atom-level feature. Second, it introduces electronic memory to extend this representation to molecules without directly computed electron density, by retrieving structurally compatible electronic environments and aligning them to the query molecular geometry. Reference selection relies only on atomic identities and nuclear geometry, without query-specific density or downstream labels, enabling reuse of precomputed electronic environments. On top of this representation, eMol uses an equivariant architecture with atom–atom, electron–electron and cross atom–electron pathways, enabling higher-order atom–electron many-body interactions to contribute to molecular representation learning. In particular, adapted five-body atom–electron couplings encode probe neighbourhoods jointly with atomic geometry, rather than directly reusing higher-order interaction formulas designed for atom-only representations. We evaluate these capabilities across quantum-chemical properties, molecular potentials, peptide-scale conformations and drug-related activity prediction, complemented by retrieval validation and interpretability analyses.

Results

Framework of eMol

Here we present eMol, an electron-density-driven molecular model that integrates electronic memory with equivariant atom–electron interactions for quantum chemical prediction (Fig. 1). The framework addresses two complementary challenges: explicitly encoding spatial electronic environments into molecular representations and transferring these environments across molecules without repeated density calculations. eMol represents a molecule by placing atoms and density-derived electron probes in a shared three-dimensional space, forming an electron-density molecular representation. Rather than simply incorporating electron density as auxiliary atomic features, eMol encodes local electronic environments as spatially positioned probes that explicitly participate in geometric interactions. This representation is constructed either from molecule-specific computed electron density, by sampling local electron probes from the density field (Fig. 1a), or from electronic memory, by retrieving a structurally compatible reference and transferring its aligned electronic environment to the query molecule without new quantum chemical density calculations (Fig. 1b). Notably, memory retrieval relies only on atomic identities and three-dimensional geometry, without requiring query electron density or downstream task labels. The resulting representation is processed by stacked (L) layers of multi-body interaction blocks, which combine atom–atom, electron–electron and atom–electron message passing with pairwise geometry and six types of higher-order three-, four- and five-body interactions (including adapted five-body atom–electron couplings that capture centered, extended and hinged biplanar configurations) to produce task-optimized atom and electronic features (Fig. 1c). Finally, we evaluate eMol across quantum-chemical prediction, molecular simulation, peptide-scale conformational modelling and drug discovery (Fig. 1d). The evaluation spans 24 quantum-chemical property-prediction tasks from QM9 and QM9-DFT, 24 molecular potential-learning systems from MD17, rMD17 and MD22, Chignolin conformational analysis and four drug-activity datasets. eMol therefore provides a unified framework for quantum-chemical learning and molecular simulation, with practical potential for electronically informed drug discovery. By decoupling electronic-environment acquisition from downstream prediction, eMol enables the reuse of electronic representations across molecular systems without repeated molecule-specific density calculations.

Quantum-chemical property prediction across molecular benchmarks

We evaluated eMol on four complementary quantum chemistry benchmarks, including MD17[35], revised MD17 (rMD17)[36], MD22[37] and QM9[38]. MD17 contains molecular dynamics trajectories of seven small organic molecules and is widely used to assess energy (kcal/mol) and force (kcal/mol/Å) prediction for learned molecular potentials. rMD17 is a revised version of MD17 with improved data consistency and corrected force labels, covering ten molecular systems. MD22 extends molecular

dynamics benchmarks to larger and more heterogeneous systems, with molecular sizes ranging from tens to hundreds of atoms and examples including peptides, nucleic-acid fragments, carbohydrates, supramolecular complexes and nanotube structures. QM9 contains more than 130,000 small organic molecules with up to nine heavy atoms and provides 12 equilibrium quantum chemical properties. These benchmarks provide a broad evaluation of eMol for both force-field learning and molecular property prediction. We compared eMol with sGDML[39], NequIP[40], Allegro[41], BOTNet[42], ViSNet[19], QuinNet[43], PaiNN-TAIP[21], Uni-Mol+[3], LiTEN[10], MGNN[25], EFA[44], and MACE[45].

We first compared eMol with representative molecular potential models on MD17 and rMD17. On MD17, eMol achieves the lowest energy MAE on five of the seven molecules and the lowest force MAE on four. For ethanol, eMol reduces the energy MAE from 0.049 with the second-best LiTEN to 0.046, corresponding to a 6.1% improvement, while reducing the force MAE from 0.059 to 0.057 (3.4%). On the larger aromatic naphthalene system, eMol achieves energy and force MAEs of 0.082 and 0.037, improving over the second-best results of 0.083 and 0.038 by 1.2% and 2.6%, respectively (Table 1). On rMD17, eMol obtains the best force MAE on five of the ten systems under the corrected force labels, with particularly pronounced gains on several molecules. For aspirin, eMol reduces the force MAE from 0.1429 with the second-best QuinNet to 0.1151, an improvement of 19.5%. Similarly, it lowers the force MAE from 0.0484 to 0.0408 for ethanol and from 0.0830 to 0.0664 for malondialdehyde, corresponding to improvements of 15.7% and 20.0%, respectively (Table 2). These results show that eMol provides accurate learned molecular potentials across both the original and corrected small-molecule trajectory benchmarks.

We next examined whether eMol extends to larger molecular systems using MD22. eMol achieves the best energy and force MAEs on Ac-Ala3-NHMe (energy, 0.0528; force, 0.0621) and AT-AT (energy, 0.0662; force, 0.0643), and the best energy MAE on DHA (0.0684) (Table 3). For Ac-Ala3-NHMe, eMol reduces the energy MAE from 0.0546 to 0.0528 and the force MAE from 0.0681 to 0.0621 relative to the second-best results, corresponding to reductions of 3.3% and 8.8%, respectively. On AT-AT, eMol achieves energy and force MAEs of 0.0662 and 0.0643, compared with the second-best values of 0.0668 and 0.0687, with the force error reduced by 6.4%. Although the performance advantage becomes less uniform for the largest systems, eMol remains competitive across diverse molecular sizes and compositions. These results indicate that eMol can extend beyond small-molecule benchmarks to systems with more heterogeneous local environments and increased structural complexity.

We further evaluated eMol on QM9 to examine its performance on equilibrium quantum chemical property prediction. eMol achieves the lowest MAE on six of the 12 properties, including α (40.5 ma_0^3), $\varepsilon_{\mathrm{HOMO}}$ (16.2 meV), $\varepsilon_{\mathrm{LUMO}}$ (14.3 meV), $\Delta\varepsilon$ (27.8 meV), ZPVE (1.05 meV) and H (4.08 meV), while remaining competitive across the other targets (Table 4). The largest relative improvement among these properties is observed for $\varepsilon_{\mathrm{HOMO}}$, where eMol reduces the MAE from 16.9 to 16.2 meV, corresponding to a 4.1% improvement over the second-best result. It also remains competitive on μ, $\langle R^2 \rangle$, U_0, U, G and C_v. Across MD17, rMD17, MD22 and QM9, eMol shows consistent advantages in learned molecular potentials and quantum chemical property prediction, supporting the effectiveness of electron-density-driven molecular representations across complementary molecular benchmarks.

Table 1: Energy and force MAE comparison on MD17. Lower values indicate better performance. Best and second-best results are shown in bold and underlined, respectively.

Molecule	Target	ViSNet	QuinNet	PaiNN-TAIP	LiTEN	MGNN	EFA	MACE	**eMol**
Aspirin	Energy	0.116	0.119	0.144	<u>0.114</u>	0.126	0.205	0.136	**0.113**
	Forces	0.155	**0.145**	0.330	0.152	0.181	0.365	0.187	<u>0.149</u>
Ethanol	Energy	0.051	0.050	0.061	<u>0.049</u>	0.057	0.060	0.051	**0.046**
	Forces	0.060	0.060	0.180	<u>0.059</u>	0.071	0.118	0.097	**0.057**
Malondialdehyde	Energy	0.075	0.078	0.091	0.074	0.083	0.095	**0.071**	<u>0.073</u>
	Forces	0.100	<u>0.097</u>	0.297	0.098	0.118	0.211	0.134	**0.095**
Naphthalene	Energy	0.085	0.101	0.093	<u>0.083</u>	0.092	0.318	0.120	**0.082**
	Forces	0.039	0.039	0.072	<u>0.038</u>	0.051	0.146	0.048	**0.037**
Salicylic acid	Energy	0.092	0.101	0.105	<u>0.091</u>	0.101	0.143	0.106	**0.090**
	Forces	<u>0.080</u>	<u>0.080</u>	0.193	**0.079**	0.096	0.223	0.108	<u>0.080</u>
Toluene	Energy	0.074	0.080	0.109	<u>0.072</u>	0.081	0.210	0.090	**0.071**
	Forces	**0.039**	**0.039**	0.090	<u>0.041</u>	0.055	0.143	0.058	0.042
Uracil	Energy	0.095	0.096	**0.090**	<u>0.091</u>	0.099	0.256	0.104	0.092
	Forces	0.062	0.062	0.130	<u>0.057</u>	0.073	0.154	0.083	**0.055**

Quantum-chemical property prediction for peptide conformations

We next tested eMol beyond standard quantum-chemical benchmarks in a more complex conformational regime. We evaluated eMol on Chignolin, a conformationally flexible peptide-scale system with 9,543 conformations and 166 atoms per conformation. To characterize the sampled conformational space, each structure was projected onto two backbone-distance coordinates,

Table 2: Energy and force MAE comparison on rMD17. Lower values indicate better performance. Best and second-best results are shown in bold and underlined, respectively.

Molecule	Target	NequIP	Allegro	BOTNet	ViSNet	QuinNet	LiTEN	MGNN	EFA	MACE	**eMol**
Aspirin	Energy	0.0530	0.0530	0.0530	0.0445	0.0486	**0.0415**	0.0715	1.0212	0.0553	0.0433
	Forces	0.1891	0.1684	0.1960	0.1520	0.1429	0.1499	0.2099	0.4123	0.2029	**0.1151**
Azobenzene	Energy	0.0161	0.0277	0.0161	0.0156	0.0394	**0.0115**	0.0530	0.9704	0.0208	0.0141
	Forces	0.0669	0.0600	0.0761	0.0585	0.0513	0.0507	0.1430	0.2749	0.0830	**0.0476**
Benzene	Energy	0.0009	0.0069	**0.0007**	**0.0007**	0.0096	**0.0007**	**0.0007**	0.4151	**0.0007**	0.0074
	Forces	0.0069	0.0046	0.0069	0.0056	0.0047	0.0046	0.0046	0.0822	0.0083	**0.0045**
Ethanol	Energy	0.0092	0.0092	0.0092	0.0078	0.0096	**0.0069**	**0.0069**	0.0499	0.0138	0.0074
	Forces	0.0646	0.0484	0.0738	0.0522	0.0516	0.0553	0.0623	0.1839	0.0992	**0.0408**
Malondialdehyde	Energy	0.0184	0.0138	0.0185	0.0132	0.0168	0.0138	0.0184	0.0780	0.0184	**0.0124**
	Forces	0.1180	0.0830	0.1338	0.0893	0.0875	0.0969	0.1176	0.2335	0.1361	**0.0664**
Naphthalene	Energy	**0.0046**	**0.0046**	**0.0046**	0.0057	0.0174	**0.0046**	0.0115	0.9581	0.0069	0.0052
	Forces	0.0300	**0.0208**	0.0415	0.0291	0.0242	0.0254	0.0600	0.1858	0.0415	0.0250
Paracetamol	Energy	0.0323	0.0346	0.0300	0.0258	0.0362	**0.0254**	0.0553	0.1389	0.0323	0.0268
	Forces	0.1361	0.1130	0.1338	0.1029	**0.0979**	0.1107	0.1568	0.2884	0.1430	0.1225
Salicylic acid	Energy	0.0161	0.0208	0.0185	0.0161	0.0330	0.0161	0.0300	0.3253	0.0208	**0.0159**
	Forces	0.0922	**0.0669**	0.0992	0.0795	0.0771	0.0945	0.1476	0.2929	0.1176	0.0754
Toluene	Energy	0.0069	0.0092	0.0069	0.0059	0.0139	**0.0046**	0.0092	0.1436	0.0069	0.0055
	Forces	0.0369	0.0415	0.0438	0.0264	**0.0244**	0.0277	0.0553	0.2082	0.0484	0.0282
Uracil	Energy	0.0092	0.0138	0.0092	**0.0069**	0.0149	**0.0069**	0.0092	0.1512	0.0115	0.0124
	Forces	0.0669	**0.0415**	0.0738	0.0495	0.0487	0.0553	0.0623	0.2322	0.0784	0.0519

Table 3: Energy and force MAE comparison on MD22. Lower values indicate better performance. Best and second-best results are shown in bold and underlined, respectively.

Molecule	Target	sGDML	ViSNet	ViSNet -Improper	QuinNet	LiTEN	MGNN	EFA	MACE	**eMol**
Ac-Ala3-NHMe	Energy	0.3910	0.0636	0.0546	0.0840	0.0597	0.0893	1.1772	0.0620	**0.0528**
	Forces	0.7900	0.0830	0.0709	0.0681	0.0763	0.1094	0.3231	0.0837	**0.0621**
AT-AT	Energy	0.7200	0.0708	0.0668	0.1440	0.1263	0.1184	0.3989	0.0913	**0.0662**
	Forces	0.6900	0.0812	0.0776	0.0687	0.0915	0.1169	0.2997	0.0981	**0.0643**
AT-AT-CG-CG	Energy	1.4200	0.1960	0.1970	0.3790	0.1743	0.2147	0.4765	**0.1556**	0.1814
	Forces	0.7000	0.1480	0.1390	0.1273	0.1490	0.1836	0.4004	**0.1168**	0.1351
DHA	Energy	1.2900	0.0741	0.0700	0.1180	0.0876	0.0968	0.5776	0.1248	**0.0684**
	Forces	0.7500	0.0598	0.0554	**0.0515**	0.0554	0.0827	0.4447	0.0604	0.0564
Buckyball catcher	Energy	1.1700	0.5080	0.5370	0.5630	**0.3440**	0.4982	0.5240	0.4624	0.5031
	Forces	0.6800	0.1840	0.2010	0.1091	0.1143	0.1827	0.4140	**0.0861**	0.1236
Stachyose	Energy	4.0000	0.0915	0.0882	0.2260	**0.0845**	0.1376	0.9115	0.1244	0.0876
	Forces	0.6800	0.0879	0.0802	**0.0543**	0.0617	0.1185	0.5547	0.0876	0.0622
Double-walled nanotube	Energy	4.0000	0.8000	**0.6010**	1.8100	0.6484	0.9735	1.7671	1.2653	0.6182
	Forces	0.5200	0.3620	0.2920	0.2473	**0.2139**	0.3864	0.7403	0.2613	0.2536

$d_{\mathrm{D3-G7}}$ and $d_{\mathrm{E5-T8}}$, defined by the D3 carbonyl O–G7 backbone N and E5 carbonyl O–T8 backbone N distances, respectively. The conformational free energy was estimated from the occupancy probability as $F = -k_{\mathrm{B}}T \ln P$ at 300 K and shifted to a zero minimum. The resulting landscape contains a compact low-free-energy basin near $d_{\mathrm{D3-G7}} \approx 3$ Å and $d_{\mathrm{E5-T8}} \approx 6.5$–$7.0$ Å, together with broader heterogeneous regions spanning larger backbone distances. This setting tests whether a model remains accurate not only around stable conformations, but also across structurally diverse peptide states. Across the six representative regions highlighted on the landscape, eMol gives the lowest local energy MAE in every case, reducing the error by 32.2–64.2% relative to the best non-eMol baseline in the corresponding region (Fig. 2a). Averaging over these highlighted regions, eMol achieves a mean regional energy MAE of 5.10 kcal mol^{-1}, compared with 9.57, 10.42, 10.92 and 12.18 kcal mol^{-1} for LiTEN, MGNN, ViSNet and QuinNet, respectively.

The error-distribution analyses show that this advantage is maintained across the full set of sampled conformations. For energy prediction, the error maps compare centered ground-truth energy with prediction error for each conformation. eMol produces a more compact error cloud centred near zero, whereas LiTEN and MGNN show broader error spreads across the sampled energy range (Fig. 2b). The probability density curves quantify this trend: eMol has a narrower energy-error distribution (-0.06 ± 6.39 kcal mol^{-1}) than LiTEN (0.17 ± 12.04 kcal mol^{-1}) and MGNN (-0.02 ± 13.04 kcal mol^{-1}) (Fig. 2c). For force prediction, the force-error maps similarly show that eMol produces a denser concentration of force-component errors around zero (Fig. 2d). The corresponding probability density curves further confirm a narrower force-error distribution for

Table 4: MAE comparison on QM9 quantum chemical property prediction. Lower values indicate better performance. Best and second-best results are shown in bold and underlined, respectively.

Target	Unit	ViSNet	QuinNet	Uni-Mol+	LiTEN	MGNN	EFA	MACE	**eMol**
μ	mD	**9.5**	771	40.1	9.8	10.0	95.9	15.7	10.8
α	ma_0^3	41.1	47	84.3	40.8	41.0	214.7	43.1	**40.5**
$\varepsilon_{\mathrm{HOMO}}$	meV	17.3	20.4	23.9	16.9	23.2	67.4	26.8	**16.2**
$\varepsilon_{\mathrm{LUMO}}$	meV	14.8	17.6	21.4	14.6	17.0	63.9	22.3	**14.3**
$\Delta\varepsilon$	meV	31.7	28.2	36.8	28.5	30.0	131.4	46.2	**27.8**
$\langle R^2 \rangle$	ma_0^2	**29.8**	194	1619	30.1	40.0	2869.5	237.0	30.4
ZPVE	meV	1.56	1.26	3.15	1.08	1.17	6.20	1.25	**1.05**
U_0	meV	4.23	7.60	36.2	4.18	**4.10**	94.84	4.96	4.29
U	meV	4.25	8.40	37.6	4.22	**4.20**	96.01	5.02	4.33
H	meV	4.52	7.80	37.5	4.15	4.10	133.21	5.64	**4.08**
G	meV	5.86	8.50	33.4	5.76	**5.70**	140.71	6.58	5.82
C_v	mcal	23	24	39.7	**22.8**	23	123	25	23

eMol (-0.01 ± 0.42 kcal mol^{-1} Å^{-1}) than for LiTEN (0.05 ± 0.54 kcal mol^{-1} Å^{-1}) and MGNN (0.04 ± 0.56 kcal mol^{-1} Å^{-1}) (Fig. 2e). These results show that eMol remains reliable beyond small-molecule benchmarks and supports peptide-scale conformational evaluation across heterogeneous regions of a molecular energy landscape.

Drug-discovery applications

We next extended the evaluation of eMol from peptide-scale conformational modelling to drug-discovery settings, spanning activity prediction and compound prioritization.

Drug-relevant activity prediction and motif attribution. We first evaluated eMol on four ChEMBL-derived activity datasets spanning three target-inhibition tasks (Mpro/3CLpro, KRAS G12D and BTK) and one metabolism-related task (CYP3A4) (**Supplementary Section 1**; **Supplementary Table 1**). Compared with two recent molecular representation learning methods (MotiL[46] and JMM[47]), eMol achieved the highest area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC) on all four datasets, with AUROC/AUPRC values of 0.854/0.783, 0.941/0.796, 0.811/0.872 and 0.842/0.390, respectively (Fig. 3a). Relative to the strongest baseline on each dataset, eMol improved AUROC by 1.6%–6.2% and AUPRC by 1.6%–84.8%, demonstrating its potential for drug discovery. We further examined whether eMol, MotiL and JMM localized mechanism-relevant chemical motifs (**Supplementary Section 2**). For Mpro/3CLpro, eMol showed the strongest enrichment for piperazine, piperidine, acrylamide and three-carbon linker motifs, with more focused attribution to the annotated regions across the three representative compounds (Fig. 3b; **Supplementary Table 2**; **Supplementary Table 3**). For CYP3A4, eMol showed the strongest enrichment for methoxy/alkoxy groups, tertiary amines, morpholine and broader metabolism-relevant motifs, with the representative compounds showing similarly clearer localization than MotiL and JMM (Fig. 3b; **Supplementary Table 4**; **Supplementary Table 3**). Overall, eMol highlighted electrophilic warheads and metabolism-relevant local electronic environments consistent with established chemical knowledge of Mpro/3CLpro inhibition and CYP3A4 metabolism.

Retrospective virtual screening for SARS-CoV-2 Mpro inhibitors. Building on these activity-prediction results, we next evaluated whether eMol could retrospectively enrich compounds with independent experimental support in DrugBank. After excluding compounds overlapping with the ChEMBL datasets, we ranked the remaining DrugBank molecules using the Mpro/3CLpro classifiers and examined literature evidence among the ten highest-ranked compounds from each model (**Supplementary Section 3**). Among these high-ranking compounds, two eMol-prioritized molecules had independent experimental support, compared with one for MotiL and none for JMM (**Supplementary Table 5**). The t-distributed stochastic neighbor embedding (t-SNE) projection visualized the screened compounds relative to the DrugBank and Mpro/3CLpro chemical spaces, with Brilacidin and Pomotrelvir highlighted among the eMol-prioritized molecules (Fig. 3c). Brilacidin ranked fifth and has shown cell-based anti-SARS-CoV-2 activity ($IC_{50} = 0.565\,\mu$M; $IC_{90} = 2.63\,\mu$M)[48]. Pomotrelvir ranked sixth and directly inhibited wild-type SARS-CoV-2 Mpro ($IC_{50} = 24$ nM; $K_i = 2.7$ nM)[49]. These results show that eMol preferentially ranked compounds with previously reported experimental evidence, supporting its utility for retrospective candidate prioritization.

Discovery of KRAS G12D-targeting compounds from existing drugs. Beyond retrospective virtual screening, we next assessed whether eMol generalized to recently reported KRAS G12D-targeting chemical drug. We curated four literature-supported compounds spanning mutation-selective inhibition, targeted degradation and pan-KRAS inhibition (**Supplementary Section 4**). At a classification threshold of 0.5, eMol assigned positive predictions to HRS-4642, RMC-9805 (zoldonrasib) and MCB-294, recovering three of the four compounds, compared with two for JMM and none for MotiL (Fig. 3d; **Supplementary**

Table 6). Notably, eMol was the only model to classify RMC-9805 as positive, whereas all three models assigned ASP3082 scores below the threshold. Overall, eMol recognized the broadest subset of these recently reported KRAS G12D-targeting compounds, demonstrating its utility for prioritizing emerging candidates with diverse pharmacological mechanisms in drug discovery.

Electronic-structure interpretability analysis

Molecular quantum properties are determined by local electronic structure, including polarity, polarization and electron delocalization. We therefore examined whether electron-density-derived probes redirect model attention towards chemically meaningful electronic environments. Using the QM9-trained model, we quantified attention distributions at three chemical levels: atoms, bonds and functional-group motifs (Fig. 4a–c, **Supplementary Tables 7–9**, **Supplementary Section 5**), selecting representative categories that contrast frequent hydrocarbon cues with electronegative, heteroatom-containing, polar, fluorinated, conjugated or ring-constrained environments. At the atom level, eMol shifts attention away from H ($\Delta_{\mathrm{rel}} = -56.5\%$) and towards O (+15.3%), N (+198.2%) and F (+215.9%) compared with the atom-only model (Fig. 4a), consistent with the role of electronegative atoms in shaping local charge distribution and polarization. This shift extends from atoms to bonds: eMol increases attention on heteroatom-containing and electronically polarized bonds, including C–N (+75.9%), C=N (+62.7%), C≡N (+104.2%), C–F (+115.2%), N–H (+65.2%), N=N (+117.2%), N–N (+120.7%) and N–O (+27.1%) (Fig. 4b). At the motif level, eMol reduces attention on hydrocarbon motifs, including methyl (−36.2%), methylene (−30.1%) and alkene (−15.1%), but increases attention on functional environments associated with lone pairs, dipoles, inductive effects or delocalized electronic structure, including amine (+58.1%), amide (+14.7%), nitrile (+95.8%), imine (+55.7%), fluorocarbon (+86.2%), aromatic ring (+5.6%) and five-membered ring (+10.3%) (Fig. 4c). These multi-scale statistics indicate that electron density reorients attention from frequency- or topology-driven structural cues towards electronic environments that are more directly connected to molecular quantum properties.

Furthermore, representative molecules show how these attention distributions appear in three-dimensional space (Fig. 4d–f). In formamide, eMol w/ ρ concentrates attention around the carbonyl/amide region, where O and N define a strongly polarized electronic environment (Fig. 4d, left). A similar pattern is observed in N-methylacetamide, where attention is enriched around the amide C–N/C=O region, consistent with the electron-density concentration and polarization of this functional group (Fig. 4e, left). In molecules containing strained or polar cyclic motifs, such as 2-methylcyclopropan-1-ol and 2-methyl-5-oxooxolane-3-carbonitrile, eMol w/ ρ assigns attention to ring-constrained, oxygen-containing and nitrile-bearing regions rather than only to atom-centred scaffold features (Fig. 4d,e, right). In fluorinated and conjugated molecules, including 2-fluoropyridin-4-ol and phenyl formate, eMol w/ ρ highlights F-containing, heteroatom-rich and carbonyl/conjugated regions that overlap with high-density or strongly polarized parts of the molecule (Fig. 4f). In contrast, eMol w/o ρ produces attention surfaces that are more diffuse or less aligned with these electron-density-rich functional regions.

We further quantified the relationship between atom-level attention and real-space electronic structure across 500 QM9-DFT test molecules. Trained eMol showed positive molecule-level Spearman correlations with both local electron density and the local electron-density gradient across all three neighbourhood definitions, with mean r_s values of 0.21–0.72 and two-sided sign-test $P \leq 1.84 \times 10^{-32}$. Compared with untrained eMol, these correlations increased by 23.5%–102.6% and remained consistently higher than those of the element-controlled model and shuffled attention (Fig. 4g, **Supplementary Section 6** and **Supplementary Table 10**). In summary, eMol exhibits strong electronic-environment interpretability: electron-density-derived probes redirect attention towards chemically meaningful atoms, bonds and motifs, while the learned attention remains quantitatively aligned with the real-space electronic structure underlying molecular quantum properties.

Ablation studies

We next dissected the contribution of electron-density information to eMol through three complementary analyses: molecule-specific density, transfer through electronic memory and architectural ablations.

Molecule-specific electron density. We first isolated the contribution of molecule-specific electron density using a retrieval-free experiment on a QM9-DFT subset. Following the DFT protocol used in EDBench, we computed electron-density fields for QM9 molecules and matched them to the original entries, yielding 105,909 uniquely matched molecules split into 95,318 training, 5,295 validation and 5,296 test molecules. This setting removes the approximation introduced by memory retrieval and alignment, allowing eMol-DFT to use density-derived electron probes from the same molecular structure. eMol-DFT achieves the best result on nine of the twelve QM9 targets and the second-best result on the remaining three (Table 5). Notably, it obtains MAEs of 41.3 ma_0^3 for α, 36.4 meV for $\Delta\varepsilon$ and 33.0 ma_0^2 for $\langle R^2 \rangle$, while remaining competitive on μ, $\varepsilon_{\mathrm{HOMO}}$ and $\varepsilon_{\mathrm{LUMO}}$. These results demonstrate the benefit of directly incorporating molecule-specific electron density into eMol for quantum chemical property prediction.

Transferability of electronic memory. We next examined whether the electronic information available from molecule-specific density can be retained when the density environment is transferred through electronic memory. Electron-density consistency was quantified using two complementary scores: local electron-density similarity, S_{loc}^{ρ}, and radial electron-density

Table 5: **QM9 property prediction with molecule-specific electron density.** MAE comparison on the QM9-DFT subset. eMol uses electron probes retrieved from electronic memory, whereas eMol-DFT uses molecule-specific DFT electron density without retrieval. Lower values indicate better performance. Best and second-best results are shown in bold and underlined, respectively.

Target	Unit	LiTEN	MGNN	QuinNet	eMol-DFT (QuiNet 5-body@I)	eMol	**eMol-DFT**
μ	mD	**10.8**	11.2	13.2	12.4	11.4	<u>11.0</u>
α	ma_0^3	43.0	43.0	50.2	44.6	<u>42.1</u>	**41.3**
$\varepsilon_{\text{HOMO}}$	meV	**19.0**	26.5	27.8	24.1	20.7	<u>20.0</u>
$\varepsilon_{\text{LUMO}}$	meV	**17.0**	20.0	21.5	21.8	18.1	<u>17.5</u>
$\Delta\varepsilon$	meV	<u>37.0</u>	38.5	42.3	41.2	37.3	**36.4**
$\langle R^2 \rangle$	ma_0^2	<u>34.0</u>	45.0	47	47.6	34.2	**33.0**
ZPVE	meV	<u>1.2</u>	1.3	1.4	1.5	<u>1.2</u>	**1.1**
U_0	meV	4.7	<u>4.6</u>	5.2	5.3	4.7	**4.5**
U	meV	4.9	<u>4.8</u>	5.4	5.5	4.9	**4.7**
H	meV	4.7	<u>4.6</u>	5.2	5.3	4.7	**4.5**
G	meV	<u>6.3</u>	<u>6.3</u>	6.9	7	6.4	**6.2**
C_v	mcal	<u>25.0</u>	<u>25.0</u>	26.5	26.4	25.2	**24.5**
IT	mol/s	655.9	**998.4**	<u>688.2</u>	45.7	27.9	46.8
Params	M	<u>4.45</u>	**1.76**	9.08	49.2		

similarity, S^{ρ}_{rad} (**Supplementary Section 7**). Across MCS, ECFP4 and MACCS similarity bins, both scores increased with structural similarity for eMol-retrieved molecular pairs (Fig. 5a; **Supplementary Table 11**). For example, across MCS bins, S^{ρ}_{loc} increased from 0.516 to 0.592, whereas S^{ρ}_{rad} increased from 0.535 to 0.764. Similar trends were observed for ECFP4 and MACCS bins, with the highest-bin S^{ρ}_{rad} reaching 0.817 and 0.832, respectively. As a control, random molecular pairs sampled within the same MCS intervals showed substantially lower electron-density similarity than eMol-retrieved pairs (**Supplementary Table 12**). In the highest MCS bin, random pairs reached $S^{\rho}_{\text{loc}} = 0.402$ and $S^{\rho}_{\text{rad}} = 0.474$, compared with 0.592 and 0.764 for eMol retrieval. These results show that eMol retrieval preserves density-derived electronic environments that are consistent with the query molecule, especially when the retrieved reference is structurally close to the query.

Correlation analysis further showed positive associations between conventional structural similarity and electron-density consistency in eMol-retrieved pairs (Fig. 5b; **Supplementary Table 13**). Pearson coefficients ranged from 0.32 to 0.38 for S^{ρ}_{loc} and from 0.58 to 0.65 for S^{ρ}_{rad}, while Spearman coefficients ranged from 0.33 to 0.39 and from 0.61 to 0.68, respectively. All reported correlations satisfied $P \leq 1 \times 10^{-7}$. Representative examples show that the density retrieved from electronic memory closely matches the density computed directly for the query molecule after alignment (Fig. 5c). The paired density maps preserve similar high-density regions and spatial density patterns around the molecular scaffold, even when the retrieved molecules are not identical to the query. In summary, these results support electronic memory as a practical route for transferring high-quality density-derived environments to molecules without new density calculations.

Architectural contributions. Finally, we assessed the contributions of the atom–electron branch, electron–electron branch, adapted 5-body interactions and probe descriptors. Removing any of these components consistently degrades performance on MD17, with the largest effect observed when the atom–electron branch is removed, increasing the average MAE by 19.6% across energy and force predictions (Table 6). Removing the adapted 5-body interactions or probe descriptors increases the average MAE by 14.5% and 11.0%, respectively, whereas removing the electron–electron branch produces a smaller but consistent increase of 7.1%. In the QM9-DFT setting, directly applying the QuiNet 5-body@I formulation performs consistently worse than the adapted 5-body interactions used in eMol-DFT across all 12 properties (Table 5), supporting the importance of adapting higher-order interactions to the heterogeneous atom–electron representation. These results show that eMol benefits from jointly modelling density-derived electron probes and their interactions with the atomic structure.

Discussion

eMol reframes electron density from a molecule-specific quantum-chemical output into a transferable representation for molecular learning. By coupling atomic scaffolds with density-derived probes, the model explicitly integrates nuclear geometry with local electronic environments associated with polarization, charge redistribution and chemical reactivity. Its consistent performance across molecular potentials, quantum chemical properties and drug-related activity prediction suggests that this representation provides a broadly useful inductive bias rather than an advantage restricted to a single task or molecular regime.

A central practical challenge is that accurate electron densities are expensive to calculate for every target molecule. Electronic memory addresses this limitation by retrieving and aligning previously computed density environments, separating the use of electronic information from the need for repeated quantum chemical calculations. The observed relationship between structural similarity and density consistency supports the premise that electronically meaningful local environments can be transferred between compatible molecules. Direct-density and memory-based experiments therefore represent complementary regimes: the former establishes the value of accurate electronic information, whereas the latter enables its scalable reuse.

Table 6: **Ablation study on MD17.** Lower values indicate better performance. Subscripts report the percentage MAE increase relative to the complete eMol model.

Molecule	Target	w/o atom-electron branch	w/o electron-electron branch	w/o adapted 5-body interactions	w/o probe descriptors	**eMol**
Aspirin	Energy	0.134 $_{\uparrow 18.6\%}$	0.121 $_{\uparrow 7.1\%}$	0.128 $_{\uparrow 13.3\%}$	0.125 $_{\uparrow 10.6\%}$	**0.113**
	Force	0.181 $_{\uparrow 21.5\%}$	0.158 $_{\uparrow 6.0\%}$	0.169 $_{\uparrow 13.4\%}$	0.163 $_{\uparrow 9.4\%}$	**0.149**
Ethanol	Energy	0.050 $_{\uparrow 8.7\%}$	0.050 $_{\uparrow 8.7\%}$	0.054 $_{\uparrow 17.4\%}$	0.052 $_{\uparrow 13.0\%}$	**0.046**
	Force	0.071 $_{\uparrow 24.6\%}$	0.061 $_{\uparrow 7.0\%}$	0.066 $_{\uparrow 15.8\%}$	0.064 $_{\uparrow 12.3\%}$	**0.057**
Malondialdehyde	Energy	0.086 $_{\uparrow 17.8\%}$	0.077 $_{\uparrow 5.5\%}$	0.082 $_{\uparrow 12.3\%}$	0.079 $_{\uparrow 8.2\%}$	**0.073**
	Force	0.112 $_{\uparrow 17.9\%}$	0.101 $_{\uparrow 6.3\%}$	0.107 $_{\uparrow 12.6\%}$	0.104 $_{\uparrow 9.5\%}$	**0.095**
Naphthalene	Energy	0.097 $_{\uparrow 18.3\%}$	0.086 $_{\uparrow 4.9\%}$	0.091 $_{\uparrow 11.0\%}$	0.089 $_{\uparrow 8.5\%}$	**0.082**
	Force	0.051 $_{\uparrow 37.8\%}$	0.041 $_{\uparrow 10.8\%}$	0.046 $_{\uparrow 24.3\%}$	0.044 $_{\uparrow 18.9\%}$	**0.037**
Salicylic acid	Energy	0.106 $_{\uparrow 17.8\%}$	0.095 $_{\uparrow 5.6\%}$	0.101 $_{\uparrow 12.2\%}$	0.098 $_{\uparrow 8.9\%}$	**0.090**
	Force	0.094 $_{\uparrow 17.5\%}$	0.084 $_{\uparrow 5.0\%}$	0.089 $_{\uparrow 11.2\%}$	0.087 $_{\uparrow 8.7\%}$	**0.080**
Toluene	Energy	0.076 $_{\uparrow 7.0\%}$	0.075 $_{\uparrow 5.6\%}$	0.079 $_{\uparrow 11.3\%}$	0.077 $_{\uparrow 8.5\%}$	**0.071**
	Force	0.051 $_{\uparrow 21.4\%}$	0.044 $_{\uparrow 4.8\%}$	0.048 $_{\uparrow 14.3\%}$	0.046 $_{\uparrow 9.5\%}$	**0.042**
Uracil	Energy	0.104 $_{\uparrow 13.0\%}$	0.097 $_{\uparrow 5.4\%}$	0.101 $_{\uparrow 9.8\%}$	0.099 $_{\uparrow 7.6\%}$	**0.092**
	Force	0.073 $_{\uparrow 32.7\%}$	0.064 $_{\uparrow 16.4\%}$	0.068 $_{\uparrow 23.6\%}$	0.066 $_{\uparrow 20.0\%}$	**0.055**

Once density probes were available in memory, retrieval contributed less than 0.04% of the combined retrieval and forward-pass latency, indicating that memory access itself does not constitute an inference bottleneck (**Supplementary Section 8**; **Supplementary Table 14**).

The interpretability and ablation results provide insight into why the representation is effective. Density-derived probes directed attention towards polarized bonds, electronegative atoms, heteroatom-containing motifs and other electronically distinctive environments. Ablation results further show that explicit atom–electron interactions are particularly important, while adapting higher-order interactions to the heterogeneous atom–electron representation provides additional gains over directly applying the QuiNet 5-body@I formulation. The agreement between attribution patterns, chemically meaningful environments and the ablation results indicates that eMol effectively uses electron-density information in molecular prediction.

Several limitations remain. The utility of electronic memory depends on the coverage and diversity of the underlying density database, while electron probes increase computational cost for very large molecular systems. Future work could address these limitations through improved retrieval, adaptive probe sampling and sparse atom–electron representations. More broadly, reusable electronic representations could open opportunities across chemistry and biology: accurate molecular potentials could enable simulations of conformational dynamics and biomolecular mechanisms; rapid quantum chemical property prediction could guide the exploration and design of molecules with desired properties; and electronically informed activity prediction could facilitate virtual screening and drug optimization. Extending these capabilities to larger molecular and chemical spaces could establish eMol as a general approach for bringing electronic information into molecular simulation, design and discovery without repeated molecule-specific electron-density calculations.

Methods

Construction of electron-density molecular representations

Construction from computed electron density. For a molecule with atomic numbers $\mathbf{Z} = \{Z_i\}_{i=1}^{N_a}$ and atomic coordinates $\mathbf{R}^a = \{\mathbf{r}_i^a\}_{i=1}^{N_a}$, eMol couples the atomic scaffold with density-derived electron probes in a shared three-dimensional coordinate system. When molecule-specific electron density is available, the density field is represented by local samples

$$\mathscr{D} = \left\{ \left(\mathbf{r}_j^{\rho}, \rho_j \right) \right\}_{j=1}^{N_\rho}, \qquad \rho_j = \rho\left(\mathbf{r}_j^{\rho} \right), \tag{1}$$

where $\rho(\mathbf{r})$, with $\mathbf{r} \in \mathbb{R}^3$, denotes the electron-density field in three-dimensional space, and $\mathbf{r}_j^{\rho}$ and ρ_j denote the coordinate and corresponding electron-density value of sample j, respectively.

$$\mathscr{G} = (\mathscr{V}_a, \mathscr{V}_e, \mathscr{E}_{aa}, \mathscr{E}_{ae}, \mathscr{E}_{ee}), \tag{2}$$

where $\mathscr{V}_a$ denotes atom nodes and $\mathscr{V}_e$ denotes density-derived electron-probe nodes, whose coordinates correspond to sampled positions in the electron-density field rather than individual electrons. $\mathscr{E}_{aa}$, $\mathscr{E}_{ae}$ and $\mathscr{E}_{ee}$ denote atom–atom, atom–electron and electron–electron relations, respectively. For nodes u and v, an edge of relation type $\ell \in \{aa, ae, ee\}$ is introduced when

$$\|\mathbf{r}_u - \mathbf{r}_v\| \leq c_\ell, \tag{3}$$

where $\mathbf{r}_u$ and $\mathbf{r}_v$ denote the three-dimensional coordinates of nodes u and v, respectively, and c_ℓ is the relation-specific cutoff radius. Unless otherwise specified, we use **eMol** to denote the memory-based setting described below, whereas **eMol-DFT** denotes the retrieval-free setting using molecule-specific DFT electron density.

Construction from electron-density memory. For molecules without molecule-specific electron density, eMol constructs the same atom–electron representation by retrieving density-derived probes from electronic memory. The memory stores reference molecules together with their electron probes and local probe descriptors. Given a query molecule, retrieval uses only atomic identities and nuclear coordinates. Memory entries are first indexed by atom count and elemental composition. The electronic memory contains approximately 2 million molecular entries, which are organized into clusters according to atom count and elemental composition to enable efficient retrieval. Exact matches are searched first; when no exact cluster is available, neighbouring clusters are ranked by normalized differences in atom count and elemental composition.

Candidate references are then ranked using structural information, including interatomic-distance distributions, bond-type distributions, element-specific coordination environments and global geometric statistics. The highest-ranked candidates are retained for atom-level correspondence using maximum common substructure matching followed by element-constrained assignment of unmatched atoms. After rigid alignment, each candidate is scored as

$$S_{\mathrm{ref}} = \mathrm{RMSD} + 0.5(1-C) + 0.5D_{\mathrm{comp}}, \qquad C = \frac{|\mathscr{M}|}{\max(N_q, N_{\mathrm{ref}})}, \tag{4}$$

where $\mathscr{M}$ is the matched atom set, N_q and N_{ref} are the numbers of atoms in the query and reference molecules, respectively, and C measures atom-match coverage. The RMSD term measures three-dimensional structural agreement after alignment, whereas D_{comp} penalizes differences in elemental composition. It is defined as

$$D_{\mathrm{comp}} = \frac{\sum_z \left| n_q(z) - n_{\mathrm{ref}}(z) \right|}{\max(N_q, N_{\mathrm{ref}})}, \tag{5}$$

where $n_q(z)$ and $n_{\mathrm{ref}}(z)$ denote the numbers of atoms of element type z in the query and reference molecules, respectively. Thus, S_{ref} favours references with close three-dimensional geometry, high atom-match coverage and similar elemental composition. The candidate with the lowest score is selected as the reference.

The selected reference molecule M_{ref} is then aligned to the query molecule M_q using matched atom pairs $\mathscr{M} = \{(i,k)\}$. The optimal rigid transformation is obtained as

$$(\mathbf{U}^*, \mathbf{t}^*) = \arg\min_{\mathbf{U}\in SO(3),\mathbf{t}} \sum_{(i,k)\in\mathscr{M}} \left\| \mathbf{r}_i^{a,q} - \left(\mathbf{U}\mathbf{r}_k^{a,\mathrm{ref}} + \mathbf{t} \right) \right\|^2. \tag{6}$$

where $\mathbf{U}$ and $\mathbf{t}$ denote the rotation matrix and translation vector, respectively. The same transformation is applied to the reference electron probes,

$$\tilde{\mathbf{r}}_m^{e,\mathrm{ref}\to q} = \mathbf{U}^* \mathbf{r}_m^{e,\mathrm{ref}} + \mathbf{t}^*, \tag{7}$$

where $\mathbf{r}_m^{e,\mathrm{ref}}$ is the coordinate of probe m in the reference molecule and $\tilde{\mathbf{r}}_m^{e,\mathrm{ref}\to q}$ is its aligned coordinate in the query frame. The local probe descriptors are preserved during this transformation.

The aligned probes are combined with the query atoms to form

$$\mathscr{G}_{\mathrm{mem}}^q = \left(\mathscr{V}_a^q, \tilde{\mathscr{V}}_e^{r\to q}, \mathscr{E}_{aa}^q, \mathscr{E}_{ae}^q, \mathscr{E}_{ee}^q \right), \tag{8}$$

where $\tilde{\mathscr{V}}_e^{r\to q}$ denotes the set of aligned electron-probe nodes whose coordinates are obtained by applying the transformation in Equation (7) to the reference probes, with their local descriptors preserved. Thus, molecule-specific density and electronic memory produce the same type of atom–electron representation, differing only in whether the probes are derived directly from the query density field or transferred from a structurally compatible reference.

Density-derived electron probes

Each atom i is initialized with a learned embedding of its atomic number,

$$\mathbf{h}_i^0 = \mathrm{Emb}(Z_i). \tag{9}$$

Electron probes provide localized representations of the surrounding electron-density environment. In the molecule-specific density setting, K probes are sampled around each atom from density samples within a local radius R_e:

$$\mathscr{P}_i = \mathrm{FPS}\left(\left\{ \mathbf{r}_j^\rho : \left\| \mathbf{r}_j^\rho - \mathbf{r}_i^a \right\| \le R_e \right\}, K \right), \tag{10}$$

where FPS$(\cdot)$ denotes farthest-point sampling[50], and $\mathscr{P}_i$ denotes the set of sampled probe coordinates around atom i. If fewer than K density samples are available within R_e, the nearest density samples are added to maintain a fixed number of probes. In the memory-based setting, the corresponding probes and descriptors are transferred from the selected reference molecule after rigid alignment.

For a probe at coordinate $\mathbf{r}_m^e$, nearby density samples are aggregated using a smooth distance-dependent kernel,

$$\omega_{mj} = \exp\left[-\frac{1}{2}\left(\frac{\left\|\mathbf{r}_m^e - \mathbf{r}_j^\rho\right\|}{\sigma_\rho}\right)^2\right] \mathrm{sigmoid}\left(\frac{\sigma_\rho - \left\|\mathbf{r}_m^e - \mathbf{r}_j^\rho\right\|}{0.35\sigma_\rho}\right), \tag{11}$$

where ω_{mj} is the contribution of density sample j to probe m, and σ_ρ controls the local density scale. The Gaussian term emphasizes nearby samples, whereas the sigmoid term smoothly suppresses long-range contributions.

The normalized kernel weights are used to construct a fixed 17-dimensional probe descriptor,

$$\phi_m \in \mathbb{R}^{17}, \tag{12}$$

which summarizes the local electron-density environment around probe m. The descriptor includes statistics of density magnitude and dispersion, density concentration, effective support size, kernel mass, radial distribution, density–distance covariance, the three-dimensional density-weighted centroid offset and higher-order density moments (**Supplementary Section 9**). These descriptors characterize both the intensity and spatial organization of the local electron density.

The probe descriptor is then projected into the model hidden space as

$$\mathbf{g}_m^0 = \mathrm{MLP}_e(\phi_m), \tag{13}$$

where MLP_e consists of a learned linear projection followed by a SiLU activation. All descriptor operations are differentiable with respect to the probe coordinates, allowing energy gradients to propagate through the electron-probe representation during force prediction.

Higher-order interactions in the atom–electron representation

Building on the higher-order geometric interactions introduced in ViSNet[19] and QuinNet[43], eMol incorporates many-body geometry into its heterogeneous atom–electron representation. The same geometric principles are applied across atom–atom, electron–electron and atom–electron relations, with additional adaptation of the 5-body interactions for atom–electron coupling (Fig. 1c).

The heterogeneous graph contains three relation types, $\mathscr{R} = \{\mathrm{AA}, \mathrm{EE}, \mathrm{AE}\}$, where AA, EE and AE denote atom–atom, electron–electron and atom–electron relations, respectively. The 3-body angular and 4-body torsional and improper interactions are applied to all three relation types, whereas the adapted 5-body interactions are used specifically for atom–electron relations.

For each relation type $\rho \in \mathscr{R}$ and edge $i \to j$, the pairwise distance and normalized direction are defined as

$$d_{ij,\rho} = \left\|\mathbf{r}_i - \mathbf{r}_j\right\|, \qquad \hat{\mathbf{r}}_{ij,\rho} = \frac{\mathbf{r}_j - \mathbf{r}_i}{d_{ij,\rho}}. \tag{14}$$

The distance $d_{ij,\rho}$ is expanded using radial basis functions, while atom-containing relations can additionally include spherical harmonic features. These pairwise quantities form the geometric basis for the higher-order interactions described below.

3-body angular interactions. For each relation type $\rho \in \mathscr{R}$, angular information is obtained by aggregating neighbouring directions around a central node. For node i,

$$\mathscr{A}_{i,\rho} = \left\| \sum_{j \in \mathscr{N}_\rho(i)} \alpha_{ij,\rho} \hat{\mathbf{r}}_{ij,\rho} \right\|^2, \tag{15}$$

which can be expanded as $\mathscr{A}_{i,\rho} = \sum_{j,k \in \mathscr{N}_\rho(i)} \alpha_{ij,\rho}\alpha_{ik,\rho} \left\langle \hat{\mathbf{r}}_{ij,\rho}, \hat{\mathbf{r}}_{ik,\rho} \right\rangle$. Here, $\mathscr{N}_\rho(i)$ denotes the neighbours of node i under relation type ρ, and $\alpha_{ij,\rho}$ is a learnable scalar coefficient derived from the corresponding edge representation. The resulting term captures angular correlations between neighbouring directions within atom–atom, electron–electron and atom–electron neighbourhoods.

The corresponding directional vector message is $\mathbf{m}_{i\to j,\rho}^{\mathrm{vec}} = a_{ij,\rho}\hat{\mathbf{r}}_{ij,\rho} + b_{ij,\rho}\mathbf{o}_{i,\rho}$, where $\mathbf{o}_{i,\rho}$ is the source vector feature in relation branch ρ, and $a_{ij,\rho}$ and $b_{ij,\rho}$ are learnable scalar coefficients. This combines the edge direction with the source vector state to propagate angular information through the corresponding relation branch.

4-body torsional interactions. For each relation type $\rho \in \mathscr{R}$, local plane normals are constructed from pairs of neighbouring directions. For node i and neighbours $j,k \in \mathscr{N}_\rho(i)$, $\mathbf{n}_{ijk,\rho} = \hat{\mathbf{r}}_{ij,\rho} \times \hat{\mathbf{r}}_{ik,\rho}$. An equivalent plane-aware representation can also

be obtained through vector rejection, $\mathrm{Rej}_{\mathbf{b}}(\mathbf{a}) = \mathbf{a} - \frac{\langle \mathbf{a},\mathbf{b}\rangle}{\|\mathbf{b}\|^2}\mathbf{b}$, which removes the component of $\mathbf{a}$ parallel to $\mathbf{b}$ and isolates the corresponding plane-aware component.

For edge $i \to j$, the torsional response couples the local planes on the source and target sides,

$$\mathscr{T}_{ij,\rho} = \left(\sum_{k \in \mathscr{N}_\rho(i)} \alpha_{ik,\rho} \mathbf{n}_{ijk,\rho} \right) \cdot \left(\sum_{l \in \mathscr{N}_\rho(j)} \beta_{jl,\rho} \mathbf{n}_{jil,\rho} \right), \tag{16}$$

which can be written as $\mathscr{T}_{ij,\rho} = \sum_{k \in \mathscr{N}_\rho(i), l \in \mathscr{N}_\rho(j)} \Gamma_{ijkl,\rho} \left\langle \mathbf{n}_{ijk,\rho}, \mathbf{n}_{jil,\rho} \right\rangle$, where $\Gamma_{ijkl,\rho}$ combines the corresponding learnable mixing coefficients. This term describes the dihedral relation between the two local planes.

4-body improper interactions. For each relation type $\rho \in \mathscr{R}$, the improper term captures out-of-plane geometry around a central node,

$$\mathscr{I}_{i,\rho} = \sum_{j \in \mathscr{N}_\rho(i)} \omega_{ij,\rho} \hat{\mathbf{r}}_{ij,\rho} \cdot \left[\left(\sum_{k \in \mathscr{N}_\rho(i)} \alpha_{ik,\rho} \hat{\mathbf{r}}_{ik,\rho} \right) \times \left(\sum_{l \in \mathscr{N}_\rho(i)} \beta_{il,\rho} \hat{\mathbf{r}}_{il,\rho} \right) \right]. \tag{17}$$

Equivalently, $\mathscr{I}_{i,\rho} = \sum_{j,k,l \in \mathscr{N}_\rho(i)} \Gamma_{ijkl,\rho} \left\langle \hat{\mathbf{r}}_{ij,\rho}, \mathbf{n}_{ikl,\rho} \right\rangle$. For $\rho = \mathrm{AA}$, EE and AE, these terms describe atomic, electronic and mixed atom–electron out-of-plane geometry, respectively.

Adapted 5-body atom–electron interactions. The 5-body interactions are applied specifically to the atom–electron branch. Directly applying the QuinNet 5-body@I formulation to the heterogeneous atom–electron representation resulted in consistently lower accuracy in the QM9-DFT control experiments (Table 5). We therefore adapted the 5-body interaction to describe higher-order couplings between atoms and density-derived electron probes. The resulting formulation contains three complementary biplanar configurations: centered, extended and hinged.

Centered biplanar interaction. The centered term couples two atom–probe planes around the same atom. For atom i,

$$\mathscr{Q}^{\mathrm{cen}}_{i,\mathrm{AE}} = \left\| \left(\sum_{p_1 \in \mathscr{N}_{\mathrm{AE}}(i)} \alpha_{ip_1} \hat{\mathbf{r}}_{ip_1,\mathrm{AE}} \right) \times \left(\sum_{p_2 \in \mathscr{N}_{\mathrm{AE}}(i)} \beta_{ip_2} \hat{\mathbf{r}}_{ip_2,\mathrm{AE}} \right) \right\|^2. \tag{18}$$

Expanding the squared norm gives

$$\mathscr{Q}^{\mathrm{cen}}_{i,\mathrm{AE}} = \sum_{p_1,p_2,p_3,p_4 \in \mathscr{N}_{\mathrm{AE}}(i)} \Gamma_{ip_1p_2p_3p_4} \left\langle \mathbf{n}_{ip_1p_2,\mathrm{AE}}, \mathbf{n}_{ip_3p_4,\mathrm{AE}} \right\rangle, \tag{19}$$

where $\mathbf{n}_{ip_ap_b,\mathrm{AE}} = \hat{\mathbf{r}}_{ip_a,\mathrm{AE}} \times \hat{\mathbf{r}}_{ip_b,\mathrm{AE}}$ denotes the normal vector of the corresponding atom–probe plane. This term measures the coupling between two such planes around a common atom.

Extended biplanar interaction. The extended term couples atom–probe planes through neighbouring atomic environments. For atom i and probes $p_1, p_2 \in \mathscr{N}_{\mathrm{AE}}(i)$, neighbouring atoms $a_1 \in \mathscr{N}_{\mathrm{EA}}(p_1)$ and $a_2 \in \mathscr{N}_{\mathrm{EA}}(p_2)$ are used to define

$$\mathscr{Q}^{\mathrm{ext}}_{i,\mathrm{AE}} = \sum_{p_1,p_2 \in \mathscr{N}_{\mathrm{AE}}(i)} \sum_{a_1 \in \mathscr{N}_{\mathrm{EA}}(p_1), a_2 \in \mathscr{N}_{\mathrm{EA}}(p_2)} \Gamma_{ip_1p_2a_1a_2} \left\langle \mathbf{n}_{ip_1a_1}, \mathbf{n}_{ip_2a_2} \right\rangle, \tag{20}$$

where $\mathbf{n}_{ip_1a_1} = \hat{\mathbf{r}}_{ip_1,\mathrm{AE}} \times \hat{\mathbf{r}}_{ia_1}$ and $\mathbf{n}_{ip_2a_2} = \hat{\mathbf{r}}_{ip_2,\mathrm{AE}} \times \hat{\mathbf{r}}_{ia_2}$. This term couples two local planes anchored at the same atom but extended through different probe-associated atomic neighbourhoods.

Hinged biplanar interaction. The hinged term couples the local plane around an atom with the local plane around a neighbouring electron probe. For an atom–electron edge $i \to p$,

$$\mathscr{Q}^{\mathrm{hinge}}_{ip,\mathrm{AE}} = \sum_{q \in \mathscr{N}_{\mathrm{AE}}(i)} \sum_{a_1,a_2 \in \mathscr{N}_{\mathrm{EA}}(p)} \Gamma_{ipqa_1a_2} \left\langle \mathbf{n}_{ipq,\mathrm{AE}}, \mathbf{n}_{pa_1a_2,\mathrm{EA}} \right\rangle, \qquad p \in \mathscr{N}_{\mathrm{AE}}(i), \tag{21}$$

where $\mathbf{n}_{ipq,\mathrm{AE}} = \hat{\mathbf{r}}_{ip,\mathrm{AE}} \times \hat{\mathbf{r}}_{iq,\mathrm{AE}}$ and $\mathbf{n}_{pa_1a_2,\mathrm{EA}} = \hat{\mathbf{r}}_{pa_1,\mathrm{EA}} \times \hat{\mathbf{r}}_{pa_2,\mathrm{EA}}$. This term therefore couples atom-side and probe-side local geometry across the atom–electron edge.

The atom–electron message is modulated by the angular, torsional, improper and 5-body responses,

$$\mathbf{m}_{p \to i,\mathrm{AE}} = \Phi_{\mathrm{AE}} \left(\mathbf{h}_i, \mathbf{g}_p, e_{ip,\mathrm{AE}}, \hat{\mathbf{r}}_{ip,\mathrm{AE}}, \mathscr{A}_{i,\mathrm{AE}}, \mathscr{T}_{ip,\mathrm{AE}}, \mathscr{I}_{i,\mathrm{AE}}, \mathscr{Q}^{\mathrm{cen}}_{i,\mathrm{AE}}, \mathscr{Q}^{\mathrm{ext}}_{i,\mathrm{AE}}, \mathscr{Q}^{\mathrm{hinge}}_{ip,\mathrm{AE}} \right), \tag{22}$$

where $\mathbf{h}_i$ and $\mathbf{g}_p$ are the atom and electron-probe features, respectively, and $e_{ip,\mathrm{AE}}$ denotes the atom–electron edge feature. The atom–atom and electron–electron branches use only the 3-body and 4-body responses,

$$\mathbf{m}_{j \to i,\rho} = \Phi_\rho \left(\mathbf{x}_i, \mathbf{x}_j, e_{ij,\rho}, \hat{\mathbf{r}}_{ij,\rho}, \mathscr{A}_{i,\rho}, \mathscr{T}_{ij,\rho}, \mathscr{I}_{i,\rho} \right), \qquad \rho \in \{\mathrm{AA}, \mathrm{EE}\}. \tag{23}$$

Thus, 3-body angular and 4-body torsional and improper interactions are shared across the three relation types, whereas the adapted 5-body terms are restricted to atom–electron interactions. This enables higher-order geometric information to be incorporated into the heterogeneous representation without explicitly enumerating high-order graph tuples.

Readout and training objective

After the final interaction block, the atomwise scalar and vector representations are passed to task-specific readout heads. For energy prediction, each atom contributes a scalar energy ε_i, and the total molecular energy is obtained by summation:

$$\hat{E} = \sum_{i=1}^{N_a} \varepsilon_i. \tag{24}$$

This atomwise aggregation yields a molecular energy that is invariant to global translation and rotation. Atomic forces are obtained as the negative gradient of the predicted energy with respect to the atomic coordinates,

$$\hat{\mathbf{F}}_i = -\frac{\partial \hat{E}}{\partial \mathbf{r}_i^a}. \tag{25}$$

Energy and force prediction are jointly optimized using

$$\mathscr{L} = \lambda_E \mathscr{L}_E + \lambda_F \mathscr{L}_F, \tag{26}$$

where $\mathscr{L}_E$ and $\mathscr{L}_F$ denote the energy and force losses, respectively, and λ_E and λ_F control their relative contributions. For molecular property prediction, the final atom representations are aggregated by a scalar property head and optimized using the corresponding target loss.

Training and evaluation

We evaluated eMol on molecular potential learning, quantum-chemical property prediction and density-informed molecular property prediction. For molecular potential benchmarks, eMol was trained jointly on molecular energies and atomic forces. QM9 evaluation covered the 12 standard quantum-chemical properties using the ViSNet data split and evaluation protocol[19]. For QM9-DFT, molecule-specific electron-density fields were computed following the DFT protocol used in EDBench[34]. After matching the density records to QM9 by elemental composition and three-dimensional structure and excluding ambiguous or duplicate matches, the resulting dataset contained 105,909 molecules, with 95,318, 5,295 and 5,296 molecules used for training, validation and testing, respectively. Model architecture, optimization and dataset-specific training settings are provided in **Supplementary Section 10**.

We further evaluated eMol as a learned molecular potential through MD17 simulations and on the Chignolin conformational landscape. Molecular dynamics simulations followed the ViSNet protocol[19], and generated trajectories were compared with reference DFT-level ab initio molecular dynamics trajectories using the interatomic-distance distribution $h(r)$. Chignolin evaluation used the same split, DFT labels and evaluation metrics as the ViSNet benchmark[19], including analysis of the two-dimensional conformational free-energy landscape. Detailed simulation and Chignolin evaluation settings are provided in **Supplementary Section 10**. Reproduction and transfer settings for LiTEN, MGNN, QuinNet, EFA and MACE are described in **Supplementary Section 11**.

Data availability

The benchmark datasets used in this study are publicly available from their original sources: MD17 [http://www.quantum-machine.org/gdml/#datasets], rMD17 [https://archive.materialscloud.org/record/2020.82], MD22 [http://www.quantum-machine.org/gdml/#datasets], QM9 [https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.datasets.QM9.html], EDBench [https://github.com/HongxinXiang/EDBench], the Chignolin dataset [https://github.com/microsoft/AI2BMD/tree/ViSNet/chignolin_data], and ChEMBL [https://www.ebi.ac.uk/chembl/]. DrugBank data used for retrospective screening were obtained under the applicable academic access terms [https://go.drugbank.com/]. MD17, rMD17, MD22 and QM9 were used following the standard splits and evaluation settings adopted in previous molecular modelling studies. The electron-density database used for retrieval follows the data construction protocol described in this work and is available through EDBench.

Code availability

All codes and the trained models are available at https://github.com/JacklinGroup/eMol.

Acknowledgements

We thank our colleagues for their helpful discussions and constructive suggestions. This research was supported by the National Natural Science Foundation of China (No. 62573372 to X.L., No. 62425204 to X.Z., No. U22A2037 to X.Z., No. 62450002 to X.Z., No. 62432011 to X.Z., No. 62672265 to T.W.), and by Southeast University Startup Grant (No. 4007032510 to D.X.), and by the Beijing Nova Program (No. 202604841280 to T.W.).

Contributions

H.X., X.L. and X.Z. conceived and designed the study. H.X. and Y.S. constructed the databases. H.X., Y.S., Q.L., J.S., T.Z. and S.J. drew the figures. H.X., Y.S. and X.L. designed the framework. Y.S., Q.L., J.S., T.Z. and H.X. developed the codes and performed experiments. H.X., Y.S., X.L., Q.L., J.S. and X.Z. performed data analysis. H.X., Y.S., Q.L., X.Z., T.W. and X.L. discussed and interpreted all results. H.X., Y.S., D.X., J.X., X.Z., T.W. and X.L. wrote and critically revised the manuscript.

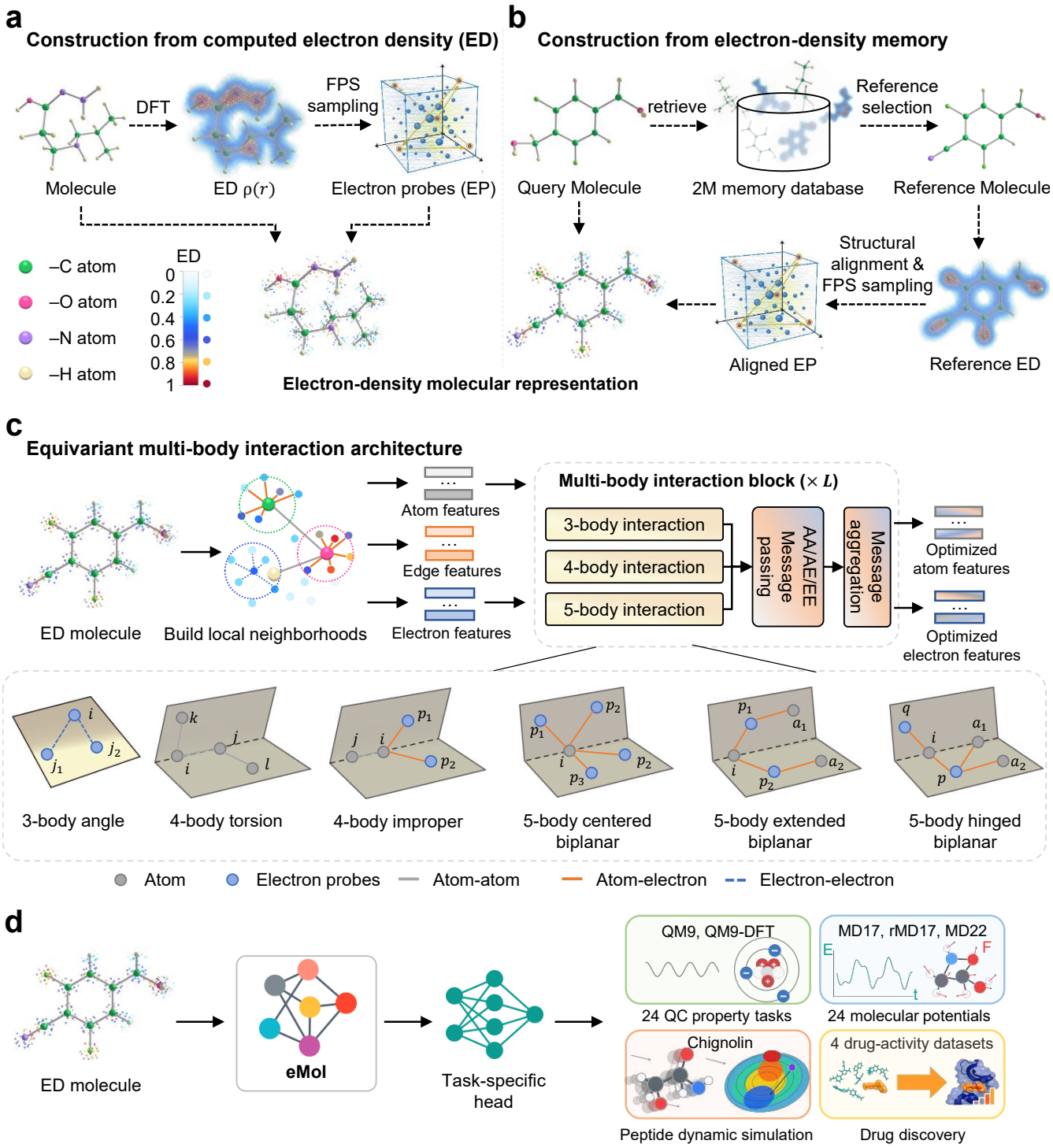


Figure 1: **Framework of eMol. a,** Construction of electron-density molecular representations from computed electron density. For molecules with molecule-specific electron density, eMol samples density-derived electron probes from the DFT-computed electron-density field and places them together with the atomic scaffold in a shared three-dimensional space to obtain an electron-density molecular representation. **b,** Construction from electron-density memory for molecules without computed electron density. eMol retrieves a structurally compatible reference from electronic memory, aligns the reference electronic environment to the query molecule and transfers its density-derived probes to construct the same electron-density molecular representation without performing new quantum chemical density calculations. **c,** Equivariant atom–electron interaction architecture. eMol embeds atoms and electron probes, constructs local atom–atom, atom–electron and electron–electron neighbourhoods, and builds pairwise and higher-order geometric features for heterogeneous interaction learning. The higher-order geometry includes 3-body angle, 4-body torsion, 4-body improper, 5-body centered biplanar, 5-body extended biplanar and 5-body hinged biplanar interactions. Stacked L layers of multi-body interaction blocks then update the representations through relation-specific atom–atom, electron–electron and atom–electron message passing and aggregation, yielding task-optimized atom features and electronic features. **d,** Training and evaluation of eMol across multiple quantum-chemical prediction, simulation and analysis tasks, including 24 quantum-chemical property prediction tasks (QM9 and QM9-DFT), 24 molecular potential-learning tasks (MD17, rMD17 and MD22), peptide-scale conformational analysis on Chignolin and four drug-activity datasets, including retrospective Mpro screening and external recovery of KRAS G12D-targeting compounds.

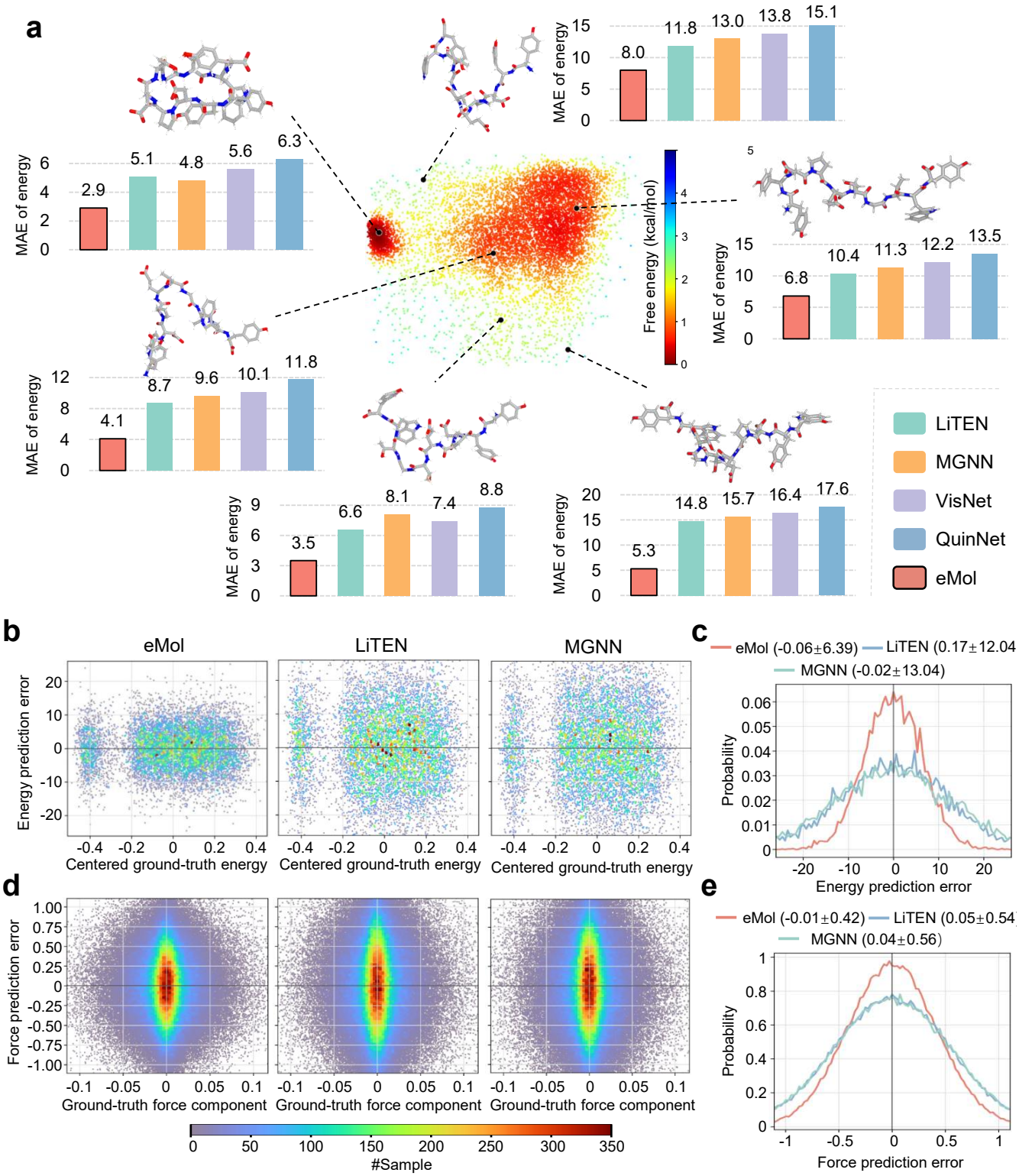


Figure 2: **Chignolin conformational landscape and prediction-error distributions. a,** Conformational free-energy landscape of Chignolin projected onto two backbone-distance coordinates: the distance between the D3 carbonyl O and G7 backbone N atoms on the x-axis, and the distance between the E5 carbonyl O and T8 backbone N atoms on the y-axis. The color map denotes the relative free energy in kcal mol^{-1}, shifted to zero at the minimum. Six representative conformational regions are highlighted and connected to their molecular structures and energy-MAE comparisons across eMol, LiTEN, MGNN, ViSNet and QuinNet. **b,** Energy prediction errors of eMol, LiTEN and MGNN, with centered ground-truth energy on the x-axis and energy prediction error on the y-axis. Each point represents one Chignolin conformation, and the color indicates local sample density. **c,** Probability density distributions of energy prediction errors for eMol, LiTEN and MGNN. Values in parentheses denote the mean and standard deviation of the error distribution. **d,** Force-component prediction errors of eMol, LiTEN and MGNN, with ground-truth force components on the x-axis and force prediction errors on the y-axis. Each point represents one force component, and the color indicates local sample density. Force errors are reported in kcal mol^{-1} Å^{-1}. **e,** Probability density distributions of force prediction errors for eMol, LiTEN and MGNN. Values in parentheses denote the mean and standard deviation of the error distribution.

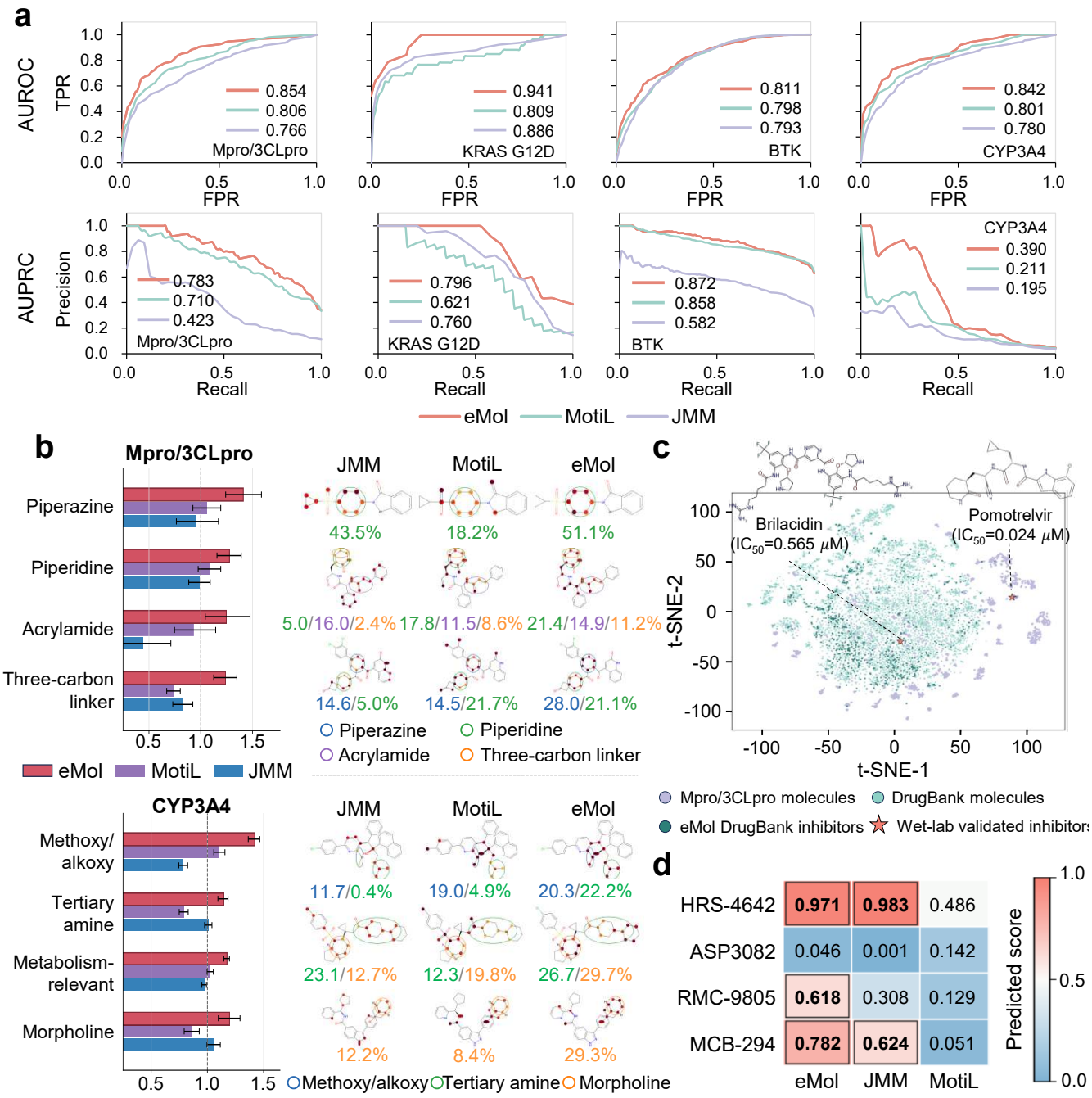


Figure 3: **Drug-discovery applications of eMol. a,** Receiver operating characteristic curves (top) and precision–recall curves (bottom) comparing eMol, MotiL and JMM on the Mpro/3CLpro, KRAS G12D, BTK and CYP3A4 tasks. Values shown in the panels denote AUROC and AUPRC, respectively. **b,** Attention enrichment of mechanism-relevant chemical motifs (left) and atom-level attribution maps for three representative compounds from each task (right) in the Mpro/3CLpro and CYP3A4 datasets. Bars show mean enrichment, error bars indicate 95% confidence intervals obtained from 5,000 bootstrap resamples, and the dashed line marks the reference value of 1.0. Coloured circles indicate the annotated regions, and percentages report the attribution assigned to these regions. **c,** t-distributed stochastic neighbor embedding (t-SNE) visualization of molecules from the Mpro/3CLpro dataset and DrugBank small molecules scored by eMol. The literature-supported candidates Brilacidin and Pomotrelvir are highlighted and annotated with their chemical structures and reported activities. High-scoring eMol candidates span multiple regions of the DrugBank chemical space, indicating the prioritization of chemically diverse candidates. **d,** Scores predicted by eMol, JMM and MotiL to four literature-supported KRAS G12D-targeting compounds. Rows represent individual compounds and columns represent the three models. The colour scale ranges from 0 to 1 and is centred at the classification threshold of 0.5, with blue and red indicating scores below and above the threshold, respectively. Black outlines mark positive predictions with scores greater than or equal to 0.5.

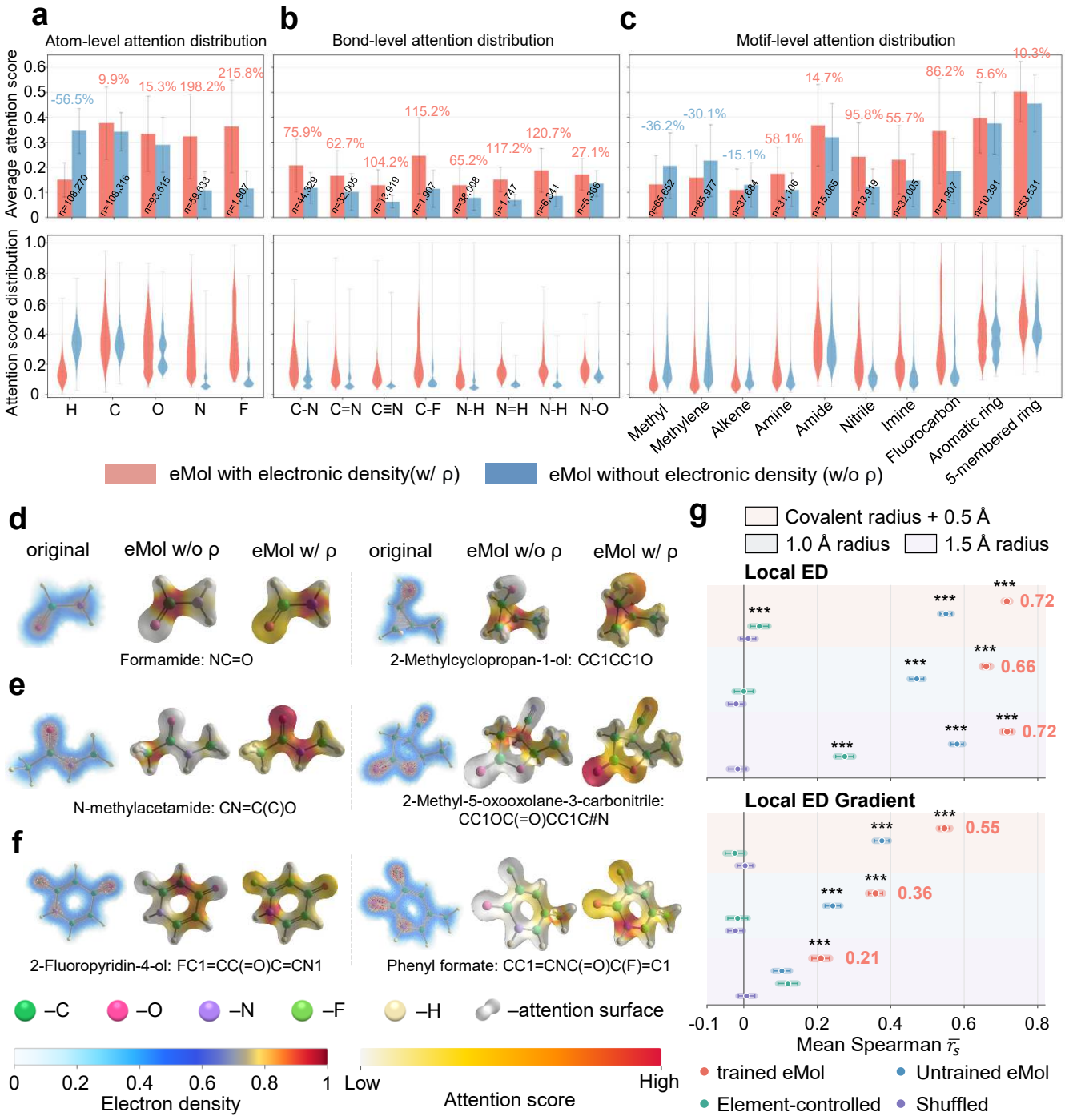


Figure 4: **Multi-scale attention analysis of eMol with and without electronic-density (ED) information. a–c,** Attention distributions of the QM9-trained eMol model with ED-derived probes (w/ ρ) and without ED-derived probes (w/o ρ) at the atom level (**a**), bond level (**b**) and motif level (**c**). Bar plots show the average attention scores, and violin plots show the corresponding attention-score distributions. Values above the bars indicate the relative change in average attention after introducing ED-derived probes, computed with respect to the w/o ρ model. The value n denotes the number of atom, bond or motif instances included in each category for attention calculation. **d–f,** Representative attention visualizations for molecules containing electronegative atoms, polar bonds and functional motifs. For each molecule, the ED distribution is shown on the left, followed by attention surfaces generated by eMol w/o ρ and eMol w/ ρ. The blue-to-red scale denotes normalized electronic density, and the yellow-to-red scale denotes attention score. Compared with the atom-only counterpart, eMol w/ ρ produces attention patterns that are more spatially aligned with electron-density-rich and chemically active regions, including heteroatom-centred environments, polarized bonds, carbonyl/amide-like groups, nitrile groups, fluorinated motifs and conjugated or ring-constrained structures. **g,** Correlations of atom-level attention with local electron density (top) and its gradient (bottom) across 500 QM9-DFT test molecules. Shaded regions denote the three neighborhood definitions: covalent radius $+0.5$ Å and fixed radii of 1.0 and 1.5 Å. Points show mean molecule-level Spearman correlations, with horizontal bars indicating 95% bootstrap confidence intervals. Colours distinguish trained eMol, untrained eMol, trained eMol (element-controlled) and shuffled attention. Asterisks indicate two-sided significance tests (sign tests) against zero ($*P < 0.05$, $***P < 0.001$). Trained eMol shows the strongest positive correlations across all settings.

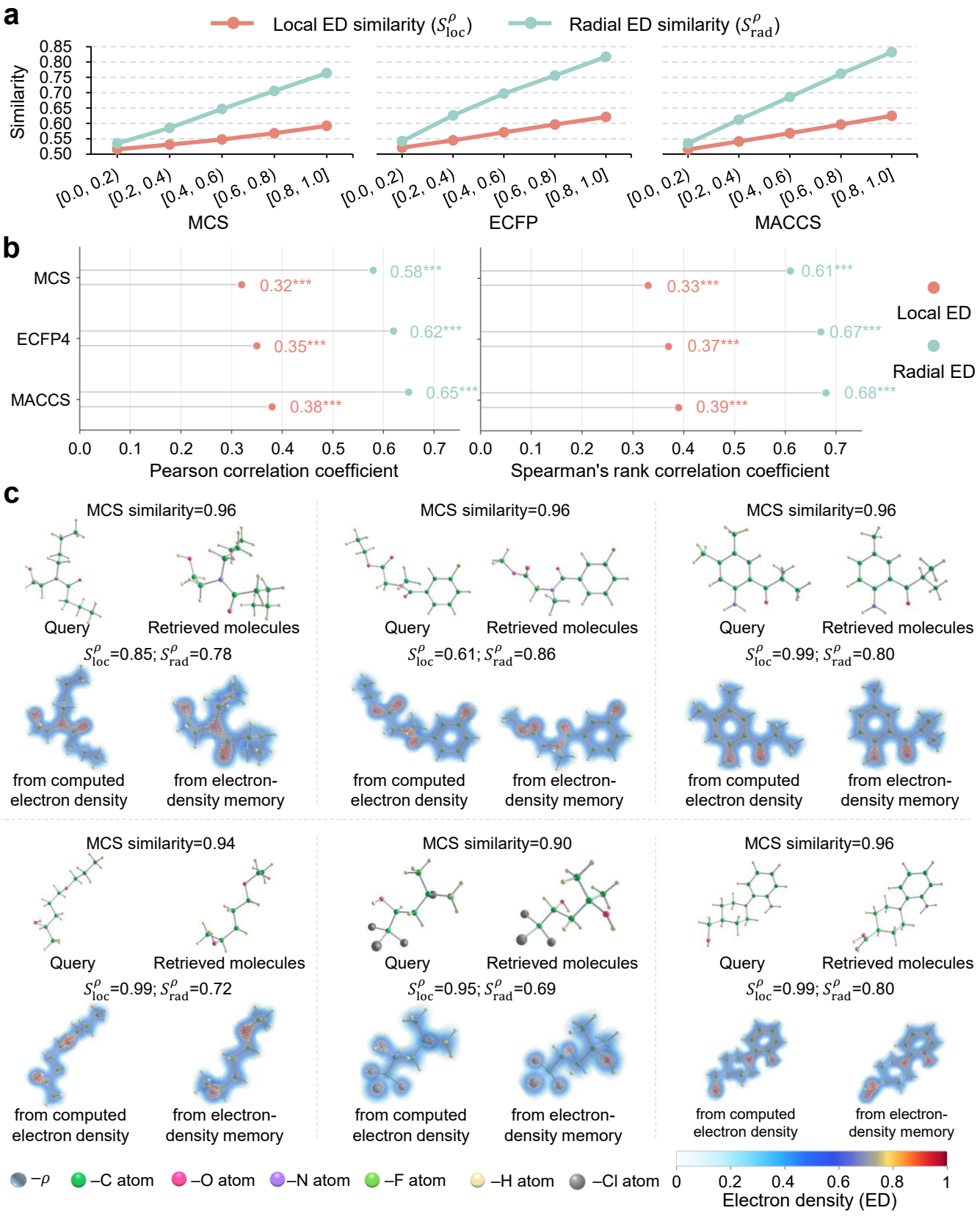


Figure 5: **Retrieval validation of the eMol electronic memory. a,** Bin-wise electron-density similarity between query molecules and eMol-retrieved molecules across MCS, ECFP4 and MACCS similarity bins. Electron-density similarity was quantified using two complementary scores: local electron-density similarity, (S^{ρ}_{loc}), and radial electron-density similarity, (S^{ρ}_{rad}). Points denote mean similarity within each bin. **b,** Pearson and Spearman correlation coefficients between conventional structural similarity metrics and electron-density similarity scores for the eMol-retrieved molecular pairs. Colours correspond to the two scores defined in a. *** denote statistical significance, with all shown values satisfying ($P = 1 \times 10^{-7}$). **c,** Representative retrieval examples from the electronic memory. For each query molecule, eMol retrieves a related molecule and its stored electron density from memory. The retrieved density is aligned to the query through MCS-guided atom correspondence and compared with the computed electron density of the query. The examples compare the density retrieved from memory with the density computed directly for the query molecule.

References

1. Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
2. Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
3. Lu, S., Gao, Z., He, D., Zhang, L. & Ke, G. Data-driven quantum chemical property prediction leveraging 3d conformations with Uni-Mol+. *Nat. Commun.* **15**, 7104 (2024).
4. Fedik, N. *et al.* Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat. Rev. Chem.* **6**, 653–672 (2022).
5. Xiang, H. *et al.* A molecular video-derived foundation model for scientific drug discovery. *Nat. Commun.* **15**, 9696 (2024).
6. Venturella, C. *et al.* Unified deep learning framework for many-body quantum chemistry via green's functions. *Nat. Comput. Sci.* **5**, 502–513 (2025).
7. Liu, Y. *et al.* An automated framework for exploring and learning potential-energy surfaces. *Nat. Commun.* **16**, 7666 (2025).
8. Han, B. & Yu, K. Refining potential energy surface through dynamical properties via differentiable molecular simulation. *Nat. Commun.* **16**, 816 (2025).
9. Gong, S. *et al.* A predictive machine learning force-field framework for liquid electrolyte development. *Nat. Mach. Intell.* **7**, 543–552 (2025).
10. Su, Q. *et al.* A scalable and quantum-accurate foundation model for biomolecular force fields via linearly tensorized quadrangle attention. *Nat. Commun.* **17** (2026).
11. Simm, G., Pinsler, R. & Hernández-Lobato, J. M. Reinforcement learning for molecular design guided by quantum mechanics. In *International Conference on Machine Learning*, 8959–8969 (2020).
12. Ajagekar, A. & You, F. Molecular design with automated quantum computing-based deep learning and optimization. *npj Comput. Mater.* **9**, 143 (2023).
13. Unke, O. T. *et al.* Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Sci. Adv.* **10**, eadn4397 (2024).
14. Kalita, B., Gokcan, H. & Isayev, O. Machine learning interatomic potentials at the centennial crossroads of quantum mechanics. *Nat. Comput. Sci.* **5**, 1120–1132 (2025).
15. Li, H. *et al.* Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nat. Comput. Sci.* **2**, 367–377 (2022).
16. Méndez-Lucio, O., Nicolaou, C. A. & Earnshaw, B. MolE: a foundation model for molecular graphs using disentangled attention. *Nat. Commun.* **15**, 9431 (2024).
17. Wang, T. *et al.* Ab initio characterization of protein molecular dynamics with AI2BMD. *Nature* **635**, 1019–1027 (2024).
18. Fan, F. *et al.* Assessing conformation validity and rationality of deep learning-generated 3D molecules. *Nat. Commun.* (2026).
19. Wang, Y. *et al.* Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nat. Commun.* **15**, 313 (2024).
20. Wang, J. *et al.* E (n)-equivariant cartesian tensor message passing interatomic potential. *Nat. Commun.* **15**, 7607 (2024).
21. Cui, T. *et al.* Online test-time adaptation for better generalization of interatomic potentials to out-of-distribution data. *Nat. Commun.* **16**, 1891 (2025).
22. Yuan, E. C.-Y. *et al.* Foundation models for atomistic simulation of chemistry and materials. *Nat. Rev. Chem.* 1–19 (2026).
23. Allen, A. E. *et al.* Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning. *npj Comput. Mater.* **10**, 154 (2024).
24. Liu, W. *et al.* Rotation-and permutation-equivariant quantum graph neural network for 3D graph data. *IEEE Transactions on Pattern Analysis Mach. Intell.* **47**, 10329–10343 (2025).
25. Chang, J. & Zhu, S. MGNN: moment graph neural network for universal molecular potentials. *npj Comput. Mater.* **11**, 55 (2025).

26. Dick, S. & Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.* **11**, 3509 (2020).

27. Huang, B., von Rudorff, G. F. & von Lilienfeld, O. A. The central role of density functional theory in the ai age. *Science* **381**, 170–175 (2023).

28. Zhang, H. *et al.* Overcoming the barrier of orbital-free density functional theory for molecular systems using deep learning. *Nat. Comput. Sci.* **4**, 210–223 (2024).

29. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864 (1964).

30. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).

31. Orio, M., Pantazis, D. A. & Neese, F. Density functional theory. *Photosynth. Res.* **102**, 443–453 (2009).

32. Schirmer, J. Review of the foundations of time-dependent density-functional theory (TDDFT). *Phys. Chem. Chem. Phys.* **27**, 4992–5005 (2025).

33. Li, W. *et al.* Predicting band gaps and band-edge positions of oxide perovskites using density functional theory and machine learning. *Phys. Rev. B* **106**, 155156 (2022).

34. Xiang, H. *et al.* EDBench: Large-scale electron density data for molecular modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems* (2026).

35. Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).

36. Christensen, A. S. & von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn. Sci. Technol.* **1**, 045018 (2020).

37. Chmiela, S. *et al.* Accurate global machine learning force fields for molecules with hundreds of atoms. *Sci. Adv.* **9**, eadf0873 (2023).

38. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).

39. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. sgdml: Constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.* **240**, 38–45 (2019).

40. Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).

41. Musaelian, A. *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).

42. Batatia, I. *et al.* The design space of e(3)-equivariant atom-centred interatomic potentials. *Nat. Mach. Intell.* **7**, 56–67 (2025).

43. Wang, Z. *et al.* Quinnet: Efficiently incorporating quintuple interactions into geometric deep learning force fields. In *Advances in Neural Information Processing Systems*, vol. 36 (2023).

44. Frank, J. T., Chmiela, S., Müller, K.-R. & Unke, O. T. Machine learning global atomic representations with euclidean fast attention. *Nat. Mach. Intell.* **8**, 388–402 (2026).

45. Kovács, D. P. *et al.* Mace-off: Short-range transferable machine learning force fields for organic molecules. *J. Am. Chem. Soc.* **147**, 17598–17611 (2025).

46. Liu, Z. *et al.* Molecular motif learning as a pretraining objective for molecular property prediction. *Nat. Commun.* (2025).

47. van Tilborg, D., Rossen, L. & Grisoni, F. Molecular deep learning at the edge of chemical space. *Nat. Mach. Intell.* 1–13 (2026).

48. Bakovic, A. *et al.* Brilacidin demonstrates inhibition of sars-cov-2 in cell culture. *Viruses* **13**, 271 (2021).

49. Tong, X. *et al.* Evaluation of in vitro antiviral activity of sars-cov-2 mpro inhibitor pomotrelvir and cross-resistance to nirmatrelvir resistance substitutions. *Antimicrob. Agents Chemother.* **67**, e00840–23 (2023).

50. Qi, C. R., Su, H., Mo, K. & Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660 (2017).

51. Komatsu, H. *et al.* Identification of sars-cov-2 main protease inhibitors from fda-approved drugs by artificial intelligence-supported activity prediction system. *J. Biomol. Struct. Dyn.* **41**, 1767–1775 (2023).

52. Cui, J. *et al.* Kras-g12d inhibitor hrs-4642 plus chemotherapy in advanced kras g12d-mutant pancreatic cancer: a phase 1b/2 trial. *Nat. Medicine* 1–8 (2026).

53. Yoshinari, T. *et al.* Discovery of kras (g12d) selective degrader asp3082. *Commun. Chem.* **8**, 254 (2025).

54. Miura, G. An assist from the chaperone: Cancer therapy. *Nat. Chem. Biol.* **21**, 1300–1300 (2025).

55. Feng, J. *et al.* A pan-kras inhibitor and its derived degrader elicit multifaceted anti-tumor efficacy in kras-driven cancers. *Cancer Cell* **43**, 1866–1884 (2025).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- nrsoftwarepolicy.pdf
- originaldataforfigures.zip
- nrreportingsummary.pdf
- SupplementaryInformation.pdf