Electronic Supplementary Information

Integrating Equivariant Architectures and Charge Supervision for Data-Efficient Molecular Property Prediction

Zixiao Yang, Hanyu Gao and Xian Kong

S1 Dataset composition and chemical coverage

Figure S1 summarizes the chemical space covered by the QM9 and QM7 datasets used for pretraining and evaluation in the main text. For each dataset, the panels show: (i) several representative molecules, (ii) histograms of heavy-atom counts resolved by element, (iii) the distribution of molecular weights, and (iv) a donut chart of functional-group frequencies. Together, these plots provide a compact view of the chemical diversity, size range and functional motifs present in the data.

For the pretraining dataset QM9, the heavy-atom statistics confirm that the vast majority of molecules are small, neutral organic compounds containing only C, N, O and F as heavy atoms. The histogram is strongly dominated by carbon, with nitrogen and oxygen appearing at lower but still substantial counts, and fluorine present as a minority substituent. The molecular-weight distribution is narrowly peaked in the $\sim$ 100–140 Da range, consistent with the intended design of QM9 as a benchmark of small fragments rather than drug-sized molecules. The functional-group donut chart shows a rich variety of chemically meaningful motifs, including aliphatic and aromatic amines, primary and secondary alcohols, ethers, carbonyl carbons(including both COO-type and other C=O groups), nitriles, epoxides, bicyclic rings and other common heteroatom-containing groups. These statistics indicate that pretraining on QM9 exposes MET to a broad set of electronic environments that resemble substructures of drug-like and materials-relevant molecules, even though each individual molecule is relatively small.

The corresponding plots for QM7 exhibit very similar behaviour. The heavy-atom histograms again show that C, N and O dominate the composition, with overall elemental ratios close to those of QM9. The molecular-weight distribution is slightly narrower and shifted towards lower masses, reflecting the fact that QM7 contains even smaller fragments on average, but the typical scale remains in the $\sim$ 80–120 Da range. The functional-group statistics reveal that QM7 shares essentially the same set of common motifs as QM9, including multiple classes of amines, alcohols, ethers, carbonyls, nitriles and small ring systems. Thus, although QM7 is used only for downstream evaluation in the main text, its chemistry is well aligned with the space on which MET was pretrained.

Taken together, these observations support the use of QM9 as a physically meaningful pretraining source for MET and justify transferring the learned representation to QM7 and other small-molecule benchmarks for test purpose. Both datasets sample a diverse but coherent region of organic chemical space, so that charge-supervised pretraining on QM9 teaches the model to distinguish between a wide range of functional environments that also occur in the downstream tasks.

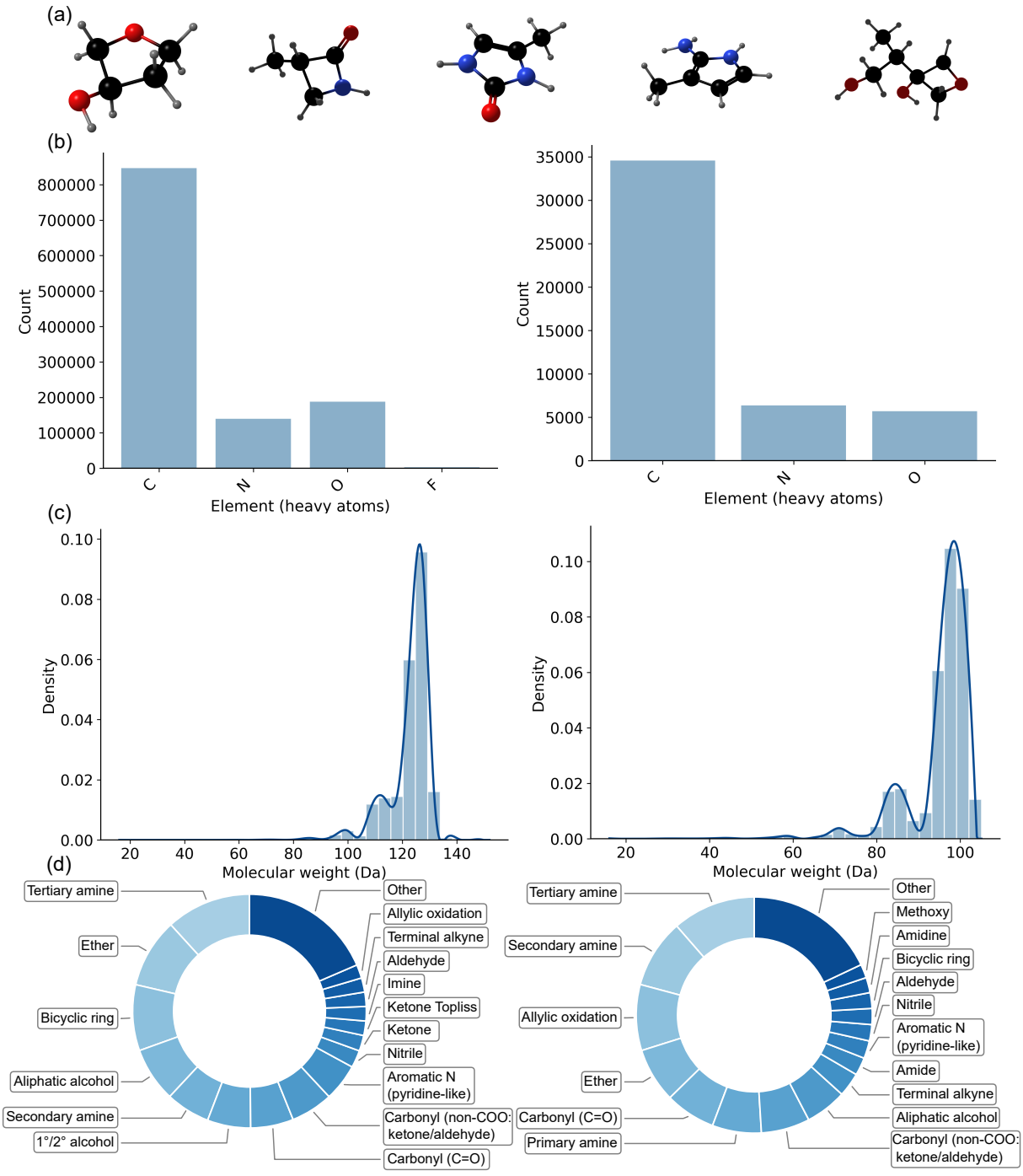

Figure S1. Summary of the chemical composition of the QM9(Left) and QM7(Right) datasets. (a) several representative molecules for QM9, (b) histograms of heavy-atom counts resolved by element, (d) the distribution of molecular weights, (d) a donut chart of functional-group frequencies. Together, these statistics illustrate that both QM9 and QM7 comprise small neutral organic molecules with similar elemental makeup and a broad variety of common functional motifs, making pretraining on QM9 well matched to downstream evaluation on QM7.