Supporting Information for Deep Learning Coordinate-free Quantum Chemistry

Matthew K. Matlock,[†] Max Hoffman,[†] Na Le Dang,[†] Dakota L. Folmsbee,[‡] Luke A. Langkamp,[‡] Geoffrey R. Hutchison,[‡,¶] Neeraj Kumar,[§] Kathryn Sarullo,[†] and S. Joshua Swamidass[*,†,‖]

[†]*Washington University in St. Louis, Department of Pathology and Immunology, Saint Louis, MO, 63130, USA*

[‡]*University of Pittsburgh, Department of Chemistry, Pittsburgh, PA, 15260, USA*

[¶]*University of Pittsburgh, Department of Chemical and Petroleum Engineering, Pittsburgh, PA, 15260, USA*

[§]*Pacific Northwest National Laboratory, Computational Biology and Bioinformatics Group, Richland, WA, 99354, USA*

[‖]*Washington University in St. Louis, Institute for Informatics, Saint Louis, MO, 63130, USA*

E-mail: swamidass@wustl.edu

List of Figures

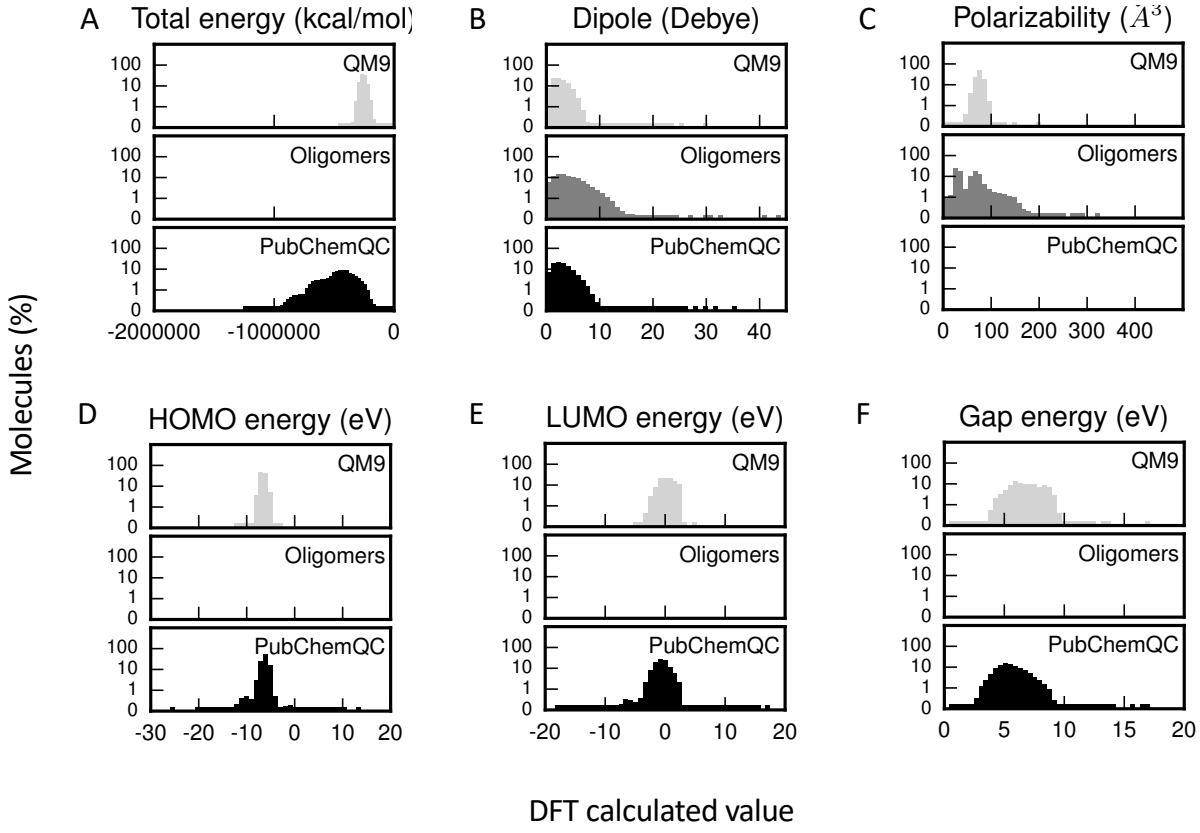

Figure S1: Datasets used in this study demonstrate a wider spectrum of quantum properties compared to QM9. (A) Distribution of total energy values in the QM9 and PubChemQC datasets (B) Distribution of dipole moments in the QM9 and PubChemQC datasets. (C) Distribution of polarizability in the QM9 and Oligomers datasets. (D) Distribution of highest occupied molecular orbital (HOMO) energy in QM9 and PubChemQC dataset. (E) Distribution of lowest unoccupied molecular orbital (LUMO) energy in QM9 and PubChemQC dataset. (F) Distribution of gap energy (difference LUMO-HOMO) in the QM9 and PubChemQC datasets.

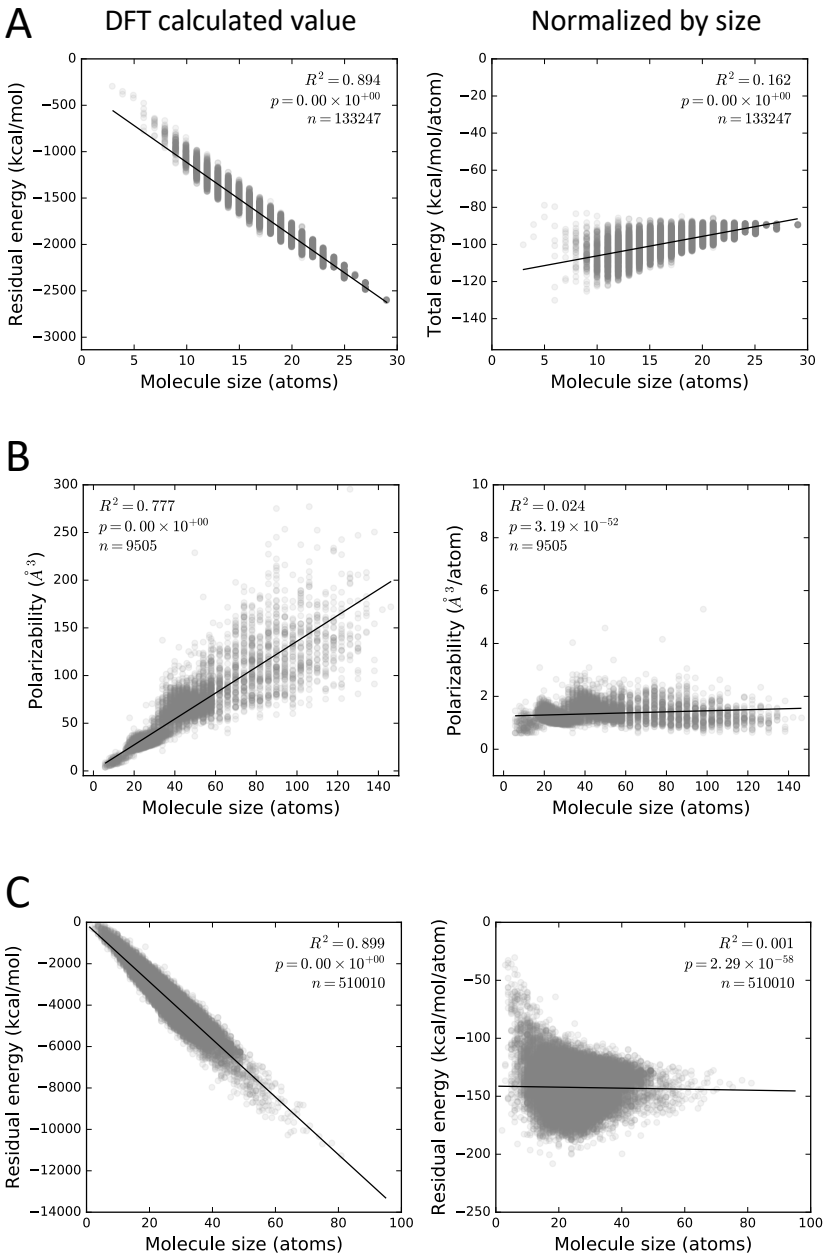

Figure S2: Several quantum measurements strongly correlated with molecule size. (A) Residual energy (difference of total energy and atom free energy) in the QM9 dataset before and after normalization. (B) Polarizability in the Oligomers dataset before and after normalization. (C) Residual energy in the PubChemQC dataset before and after normalization.

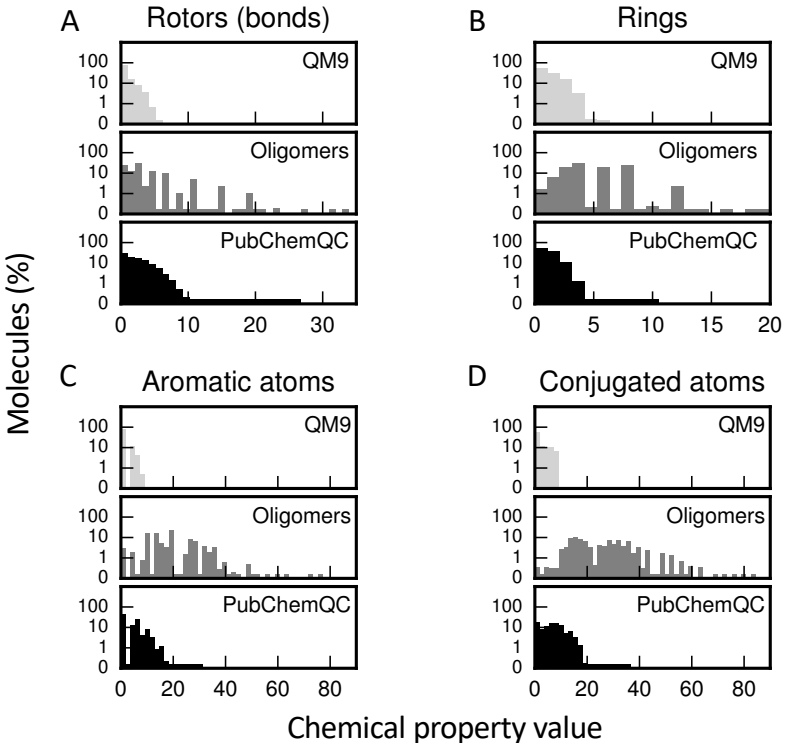

Figure S3: Datasets used in this study demonstrate a wider spectrum of chemical structures compared to QM9. (A) Distribution of rotable bond counts. (B) Distribution of ring counts. (C) Distribution of aromatic system sizes, defined as the number of atoms with at least one aromatic bond as labeled by RDKit. (D) Distribution of conjugated system sizes, defined as the number of atoms with at least one conjugated bond as labeled by RDKit.

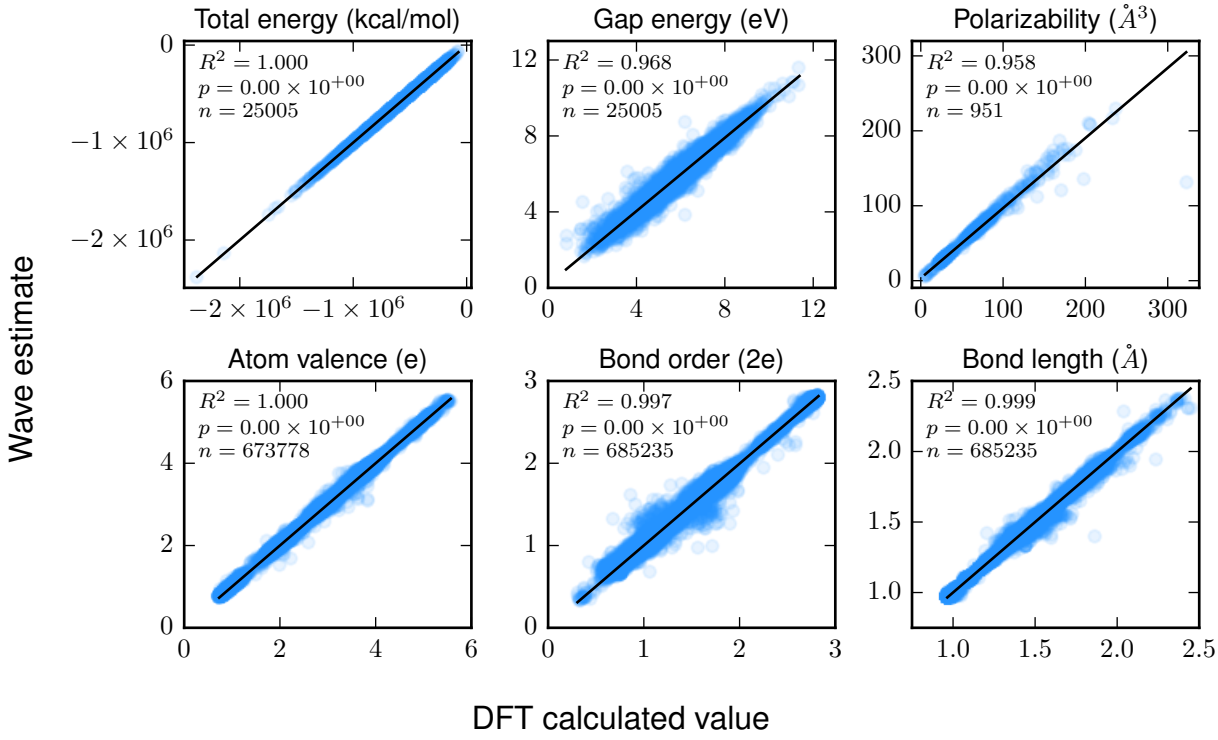

Figure S4: Coordinate free estimates of quantum properties with Wave networks are strongly correlated with DFT estimates. Scatter plots show estimates for molecules, atoms and bonds in the holdout test sets. Quality of fit (R^2) was estimated by pearson correlation.

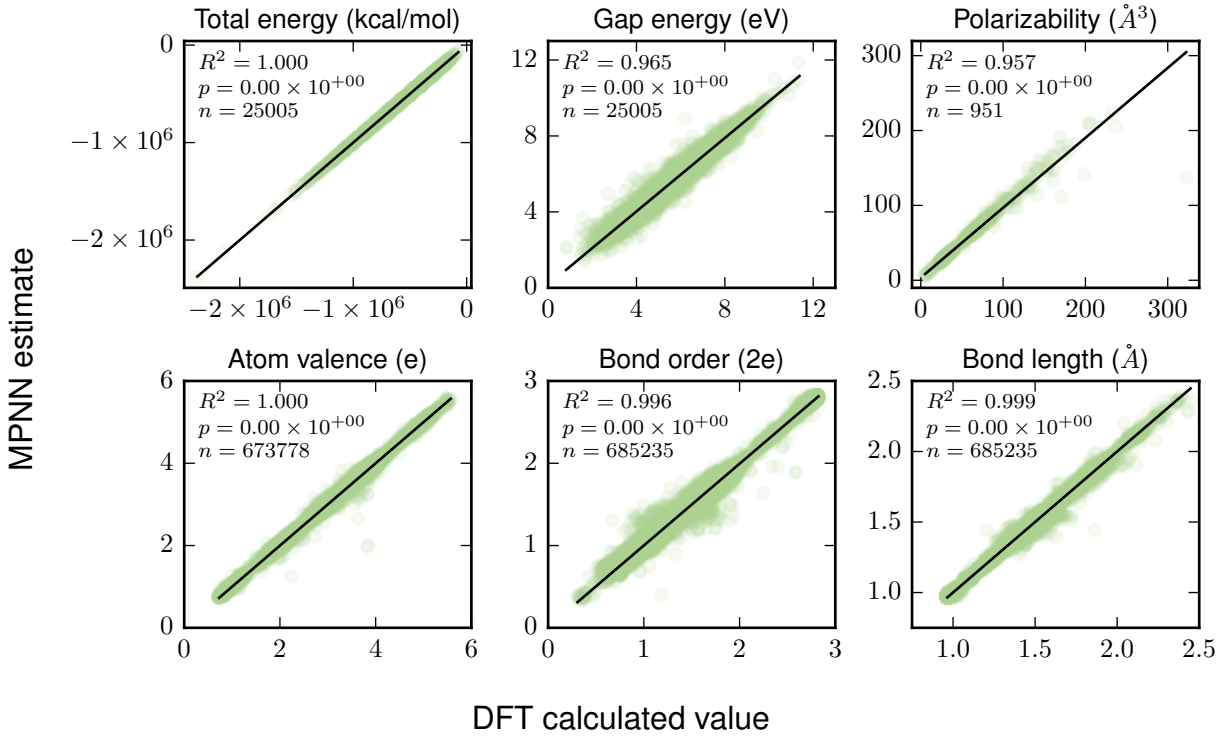

Figure S5: Coordinate free estimates of quantum properties with MPNN networks are strongly correlated with DFT estimates. Scatter plots show estimates for molecules, atoms and bonds in the holdout test sets. Quality of fit (R^2) was estimated by pearson correlation.

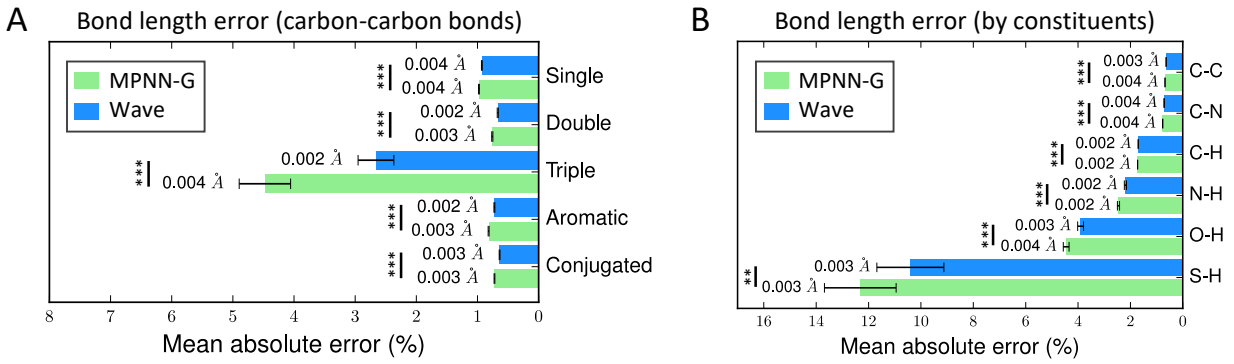

Figure S6: MPNN-G exhibits larger errors on all bond types by order and constituent atoms (A) Accuracy on carbon-carbon bond lengths by bond type. Bonds are labeled single, double or triple by kekulization. Categories single, double, aromatic and conjugated may overlap. (B) Accuracy on bond length by constituent atom pairs. Statistical tests were performed by paired t-test. **: $p < 0.01$, ***: $p < 0.001$.

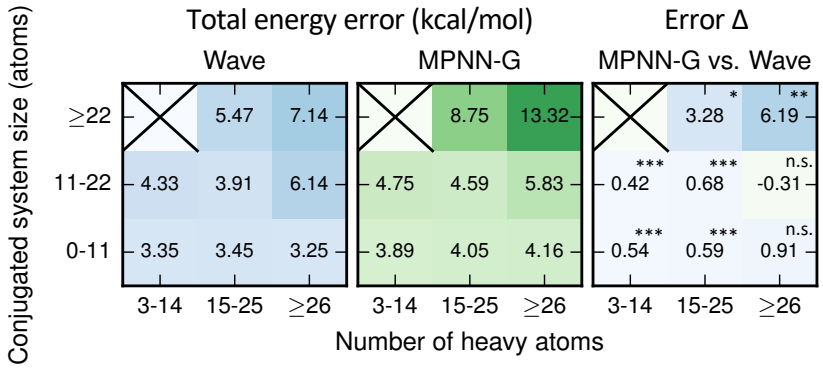

Figure S7: MPNN-G has larger absolute errors on the total energy task for both large molecules and large conjugated systems when compared to Wave. (left) Mean absolute error of Wave on the PubChemQC total energy task. (middle) Mean absolute error of MPNN-G. (right) Mean difference in absolute error between Wave and MPNN-G. Statistical tests were performed by paired t-test. ns: not significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

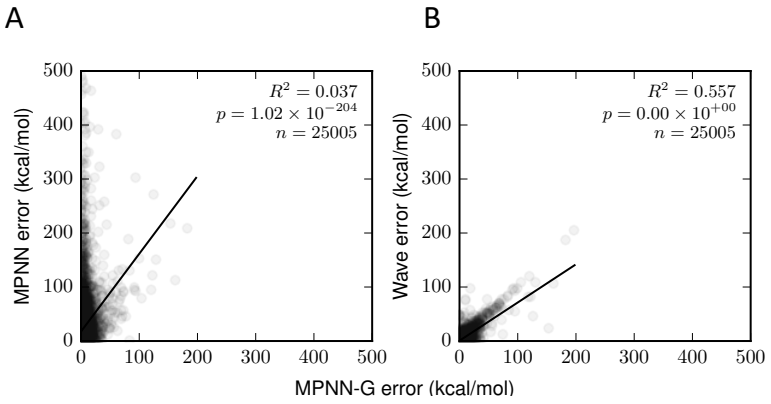

Figure S8: MPNN without global variables exhibits substantially different types of errors compared to MPNN-G or Wave. (A) The correlation of errors on total energy between MPNN and MPNN-G is low, suggesting substantially different model behavior of MPNN. (B) Total energy error of MPNN-G and Wave is strongly correlated, suggesting that these models predict similar values for similar molecules and commit the same types of errors.