Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2022

△-Quantum machine-learning for medicinal chemistry

Kenneth Atz*[1], Clemens Isert*[1], Markus N. A. Böcker[1], José Jiménez-Luna†[1, 2], and Gisbert Schneider†[1, 3]

[1]ETH Zurich, Department of Chemistry and Applied Biosciences, RETHINK, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.
[2]Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Straße 65, 88397 Biberach an der Riss, Germany
[3]ETH Singapore SEC Ltd., 1 CREATE Way, #06-01 CREATE Tower, Singapore 138602, Singapore

Contents

*Contributed equally
†Correspondence: joluna@ethz.ch, gisbert@ethz.ch

1 Training details

The `DataParallel` class implemented in the PyTorch Geometric Python (version 1.7.2) package was used to train the model in parallel on multiple graphical processing units (GPUs). The dataset splitting procedure is schematically depicted in Fig. S1. Models with training set sizes of 100k samples or more were parallelized across 4 GPUs (Nvidia GeForce GTX 1080) and trained for 72 to 120 hours, those with training set sizes of 10k were trained on a single GPU for 24h, and models for training sets containing fewer than 1k samples were trained for 4h. All models use a batch size of 16 samples. All non-production (*i.e.*, those with fewer than ∼2M training samples) models were optimized using the Adam stochastic gradient descent optimizer[1], with a starting learning rate of 10^{-4} (5×10^{-4} for the single-task orbital energies and the multi-task models), a mean squared error loss (MSE) on the training set, a decay factor of 0.7, a decay patience of 20 gradient updates for the mean absolute error (MAE) observed on the validation set, and an exponential smoothing factor of 0.9. Optimized models were only stored if they achieved an improved MAE on the chosen validation set (early stopping). The constants for the adapted loss functions (β for the formation energy loss $\mathcal{L}_{\text{form}}$ and λ for Wiberg bond order loss $\mathcal{L}_{\text{wbo}}$) were chosen from screening the following hyperparameters: $\beta \in \{0.1, 0.25, 0.5, 1, 2, 4, 8, 16\}$ and $\lambda \in \{5\times10^{-3}, 10^{-2}, 5\times10^{-2}, 10^{-1}, 5\times10^{-1}, 1\}$. Production models were trained for a fixed number of 50 total epochs for all endpoints, with an otherwise identical training setup. All models considered in this study were trained using the Leonhard and Euler clusters at ETH Zurich. Models for different training set sizes and single-task models and models for the QM9 dataset are available at `https://doi.org/10.3929/ethz-b-000520281`. A training tutorial is included in the project's GitHub repository ((`https://github.com/josejimenezluna/delfta`)). Nested training sets were used for the learning curves, so that all molecules used for training with a smaller subset were also used for training with a larger subset. This approach allowed us to study the effect of adding a certain number of molecules to the training set and its impact on model accuracy.

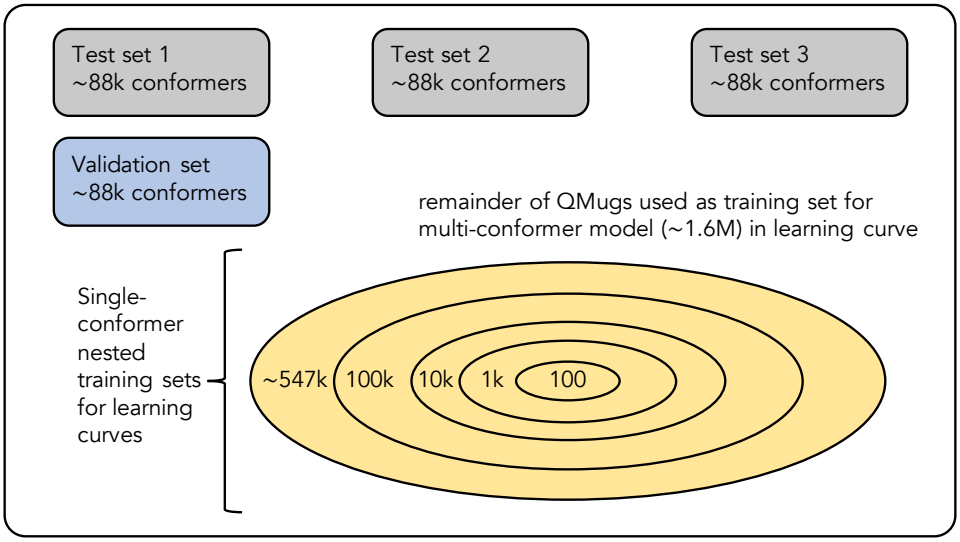


Fig. S1. Schematic of the data splitting procedure. Note that for the formation energy models, all conformers of the same molecule were grouped within the same split, yielding training set sizes of approximately 300, 3k, 30k, 300k, and 1.6M samples. For all other models a single conformer was used within the same split, yielding training set sizes of 100, 1k, 10k, 100k and 547k, 1.6M samples.

2 Scatter plots for direct-learning models

Fig. S2 shows the distribution of the direct-learning predictions relative to ground-truth reference values for molecules from the test sets.

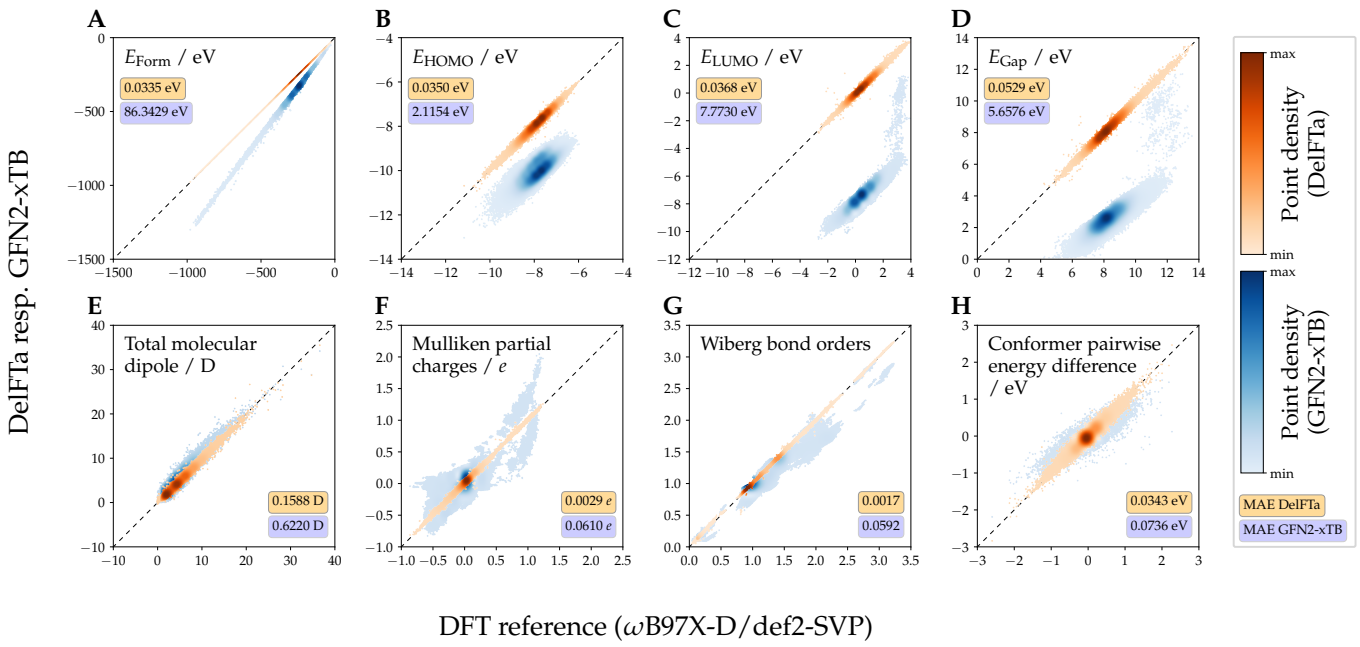


Fig. S2. Scatter plots illustrating the accuracy of the predictions provided by the trained direct-learning models and the GFN2-xTB baseline, w.r.t. DFT reference properties (ωB97X-D/def2-SVP) for ∼88k test set molecules (∼263k conformers). Direct-learning predictions obtained using the models trained on the 1.6M conformer training set, and with single-/multi-task settings as described in the Methods section. Wiberg bond order results only for bonds where a GFN2-xTB value is available (excl. two interactions for which the DFT-value is higher than 0.05). Colorbars scaled individually for each property.

3 Benchmark against ωB97X-D/def2-QZVP

We put the performance of the models trained on 1.6M conformations (both for Δ- and direct-learning paradigms) in the context of the same chosen reference functional but with a more comprehensive basis set, namely ωB97X-D/def2-QZVP[2,3]. In order to do so, for 1k randomly-sampled molecules from the considered test sets, with three conformers each, the QM properties investigated in this study were recomputed using the aforementioned basis set. Computations were performed using Psi4[4] (version 1.3.2). 958 molecules (corresponding to 2,874 conformations) for which all three conformers could be successfully computed using empirically-determined limits of computational resources (4 CPU cores for 24 h wall-time, up to 40 GB of system memory, and 400 GB of local disk space) were included in this benchmark (see `https://doi.org/10.3929/ethz-b-000520329`). MAEs w.r.t. these results were compared for the four considered methods (ωB97X-D/def2-SVP[2,3], GFN2-xTB[5–8], and the trained models in both Δ- and direct-learning settings) for all eight properties considered in this study and results are shown in Table S1.

Δ-learning models offer advantage over their directly-trained analogues for some of the endpoints considered, namely formation energies, dipoles, and conformer pairwise energy differences. Interestingly, for some others (*e.g.*, LUMO energies and HOMO-LUMO gap) the contrary holds true with regards to this basis set choice.

The machine-learning predictions are closer to ωB97X-D/def2-QZVP reference values than its baseline GFN2-xTB for the formation energy, orbital energies incl. HOMO-LUMO gap, molecular dipole, Wiberg bond orders, and conformer pairwise energy differences endpoints. However, the greater accuracy of GFN2-xTB compared to ωB97X-D/def2-SVP for Mulliken partial charges imply that neither Δ- nor directly-predicted values can improve on accuracy over the baseline.

Table S1: Benchmark results showing MAEs w.r.t. ωB97X-D/def2-QZVP reference values for 958 molecules (2,874 conformers) from the test sets. Wiberg bond order results only for covalent bonds. Bold numbers highlight the best performing method between GFN2-xTB, Δ- and direct-learning.

Property	**Unit**	**ωB97X-D/ def2-SVP**	**GFN2-xTB**	**DelFTa**	
				Δ	direct
Formation energy	eV	2.9093	85.1303	**2.9085**	2.9135
HOMO energy	eV	0.0606	2.0579	0.0756	**0.0745**
LUMO energy	eV	0.1416	7.6384	0.1404	**0.1272**
HOMO-LUMO gap	eV	0.0874	5.5807	0.0896	**0.0793**
Dipole	D	0.1203	0.6059	**0.1556**	0.1955
Mulliken partial charges	e	0.0739	**0.0609**	0.0742	0.0742
Wiberg bond orders	-	0.1006	0.1583	0.1007	**0.1004**
Conformer pairwise energy difference	eV	0.0382	0.0707	**0.0434**	0.0515

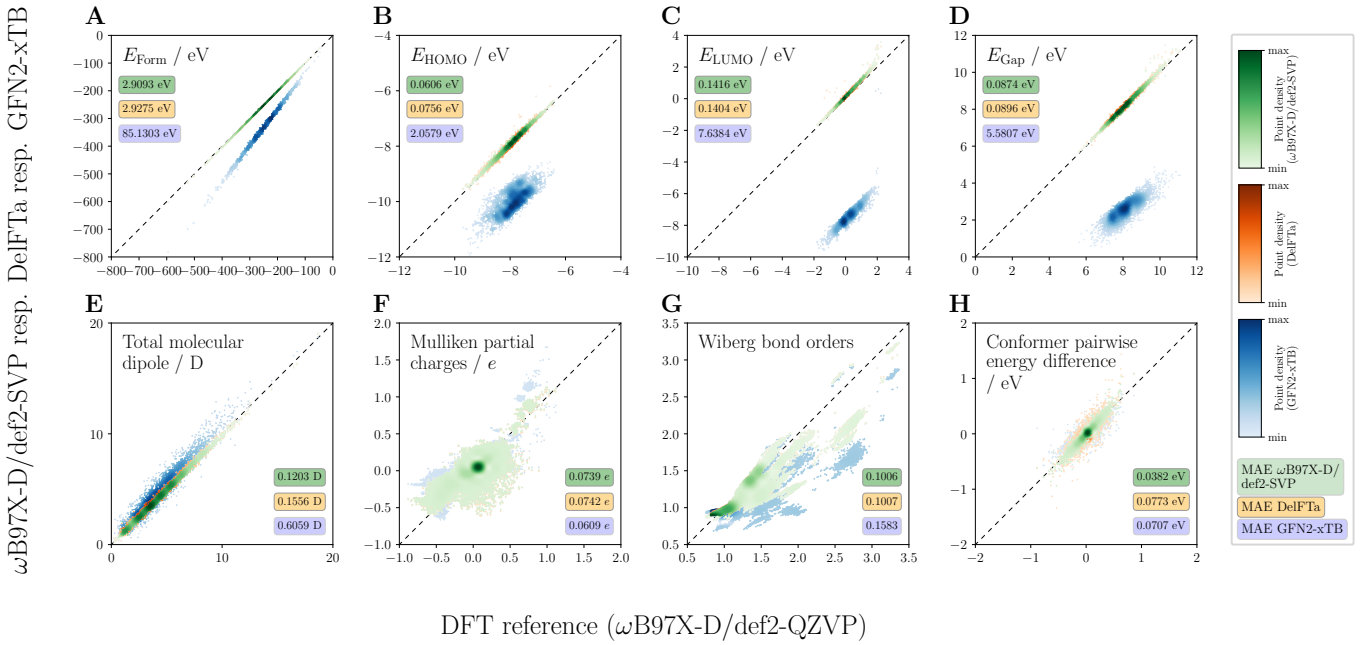


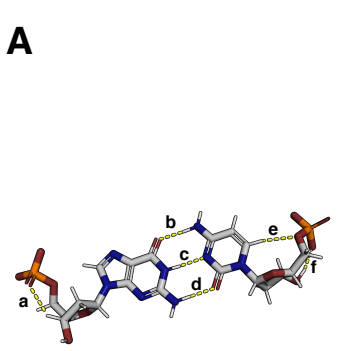
Fig. S3. Scatter plots illustrating the accuracy of the predictions provided by the trained Δ-learning models and the GFN2-xTB baseline as well as the ground truth ωB97X-D/def2-SVP computations, w.r.t. reference values computed with a larger basis set (ωB97X-D/def2-QZVP) for 958 test set molecules (corresponding to 2,874 conformations). Δ-learning predictions obtained using the models trained on the 1.6M conformer training set, and with single-/multi-task settings as described in the Methods section. Wiberg bond orders for covalent bonds only. Colorbars scaled individually for each property.

4 Non-covalent interactions in biomolecules

The structures of eight selected biomolecules extracted from the Protein Data Bank (PDB)[9] were prepared with Molecular Operating Environment (version 2019.0102). The structures were prepared using the QuickPrep module with the following parameters: `Preserve Sequence and Neutralize`, `Use Protonate 3D for Protonation: True`, `Allow ASN/GLN/HIS "Flips" in Protonate 3D: True`, `Delete: No deletions`, `Tether Receptor: No changes`, `Fix: All atoms fixed`, `Refine: No refinements`. Subsequently the structures were manually curated, whereby all atoms were removed which were farther away than one additional residue from the non-covalent interactions of interest. The resulting radicals which were generated due to the broken covalent bonds were padded with hydrogen atoms. The final number of atoms per biomolecule were in the range of 54 (PDB ID: 5GNJ) to 375 (PDB ID: 3H0O). Structures are available at `https://doi.org/10.3929/ethz-b-000520281`.

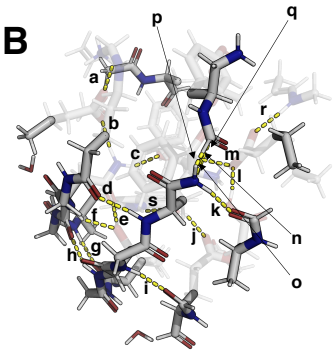
5 Reference calculations for charged molecules

To investigate the performance of the DelFTa application on charged molecules, 100 molecules were randomly chosen from the test sets and their SMILES notation was extracted. The molecules' protonation states were modified (to pH



PDB ID: 1BNA

	ωB97X-D	ML
a	0.0801	0.0060
b	0.1561	0.1343
c	0.1513	0.1082
d	0.1483	0.1396
e	0.0516	0.0053
f	0.2727	0.2057



PDB ID: 3H0O

	ωB97X-D	ML
a	0.0647	0.0705
b	0.1053	0.0987
c	0.1188	0.1154
d	0.1161	0.1141
e	0.1721	0.1693
f	0.0823	0.0725
g	0.1036	0.0908
h	0.0808	0.0765
i	0.0958	0.0888
j	0.1735	0.1632
k	0.1365	0.1302
l	0.2937	0.2537
m	0.0524	0.0549
n	0.1326	0.0735
o	0.1768	0.2605
p	0.2472	0.7140
q	0.1078	0.1065
r	0.1169	0.1075
s	0.0625	0.0618

Fig. S4. Non-covalent interactions in selected biomolecules. For panel B, backbone shown semi-transparently for visual clarity. Interactions with DFT-calculated Wiberg bond orders between 0.05 and 0.8 shown, and both calculated (ωB97X-D/def2-SVP) and predicted (ML, using models trained on 1.6M datapoints) values tabulated. Only interactions with calculated values between 0.05 and 0.8 are displayed. Interactions with solvent atoms and between atoms fewer than six covalent bonds apart not shown.

0 and 14, respectively) using Openbabel[10] (version 3.1.1). 59 molecules which exhibited a net charge different from zero were kept and processed in the same way as was done for the molecules in the QMugs data collection (generation of three conformers per molecule (totalling 177 conformers), geometry optimization, and ωB97X-D/def2-SVP QM calculations). For a detailed description of the process, see reference 11. QM calculations for 176 of the 177 conformers converged in 100 SCF iterations, the unconverged structure was discarded. The results from those calculations are available at `https://doi.org/10.3929/ethz-b-000520329`. Table S2 shows the distribution of formal charges. The machine-learning models presented in this work (trained on 1.6M molecules) were used to predict the eight endpoints. Table S3 shows the results from these calculations.

Table S2: Distribution of formal charges for 176 conformers.

Formal charge	Number of conformers
+2	24
+1	93
-1	54
-2	3
-3	2

Table S3: Benchmark results showing MAEs w.r.t. ωB97X-D/def2-SVP reference values for 176 conformers from the reference calculations for charged molecules. Wiberg bond order results only for bonds where a GFN2-xTB value is available. Bold numbers highlight the method with the lowest MAE w.r.t. reference values.

Property	**Unit**	**GFN2-xTB**	**DelFTa**	
			Δ-learning	direct-learning
Formation energy	eV	90.0	**5.90**	576
HOMO energy	eV	**0.797**	9.37	26.6
LUMO energy	eV	**0.430**	9.16	59.8
HOMO-LUMO gap	eV	**0.813**	14.2	54.9
Total molecular dipole	D	**2.22**	19.8	132
Mulliken partial charges	e	0.0588	**0.0098**	0.0240
Wiberg bond orders	-	0.0569	**0.0037**	0.0051
Conformer pairwise energy difference	eV	0.178	**0.128**	0.578

References

1. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2014).
2. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10,** 6615–6620 (2008).
3. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7,** 3297–3305 (2005).
4. Smith, D. G. *et al.* PSI4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **152,** 184108 (2020).
5. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1–86). *J. Chem. Theory Comput.* **13,** 1989–2009 (2017).
6. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15,** 1652–1671 (2019).
7. Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *J. Chem. Theory Comput.* **15,** 2847–2862 (2019).
8. Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.,* e01493 (2020).
9. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
10. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3,** 1–14 (2011).
11. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs: Quantum Mechanical Properties of Drug-like Molecules. *arXiv:2107.00367* (2021).