AlChemy: A Quantum Chemistry Dataset for Benchmarking AI Models

Guangyong Chen[1]*, **Chang-Yu (Kim) Hsieh**[1], **Chee-Kong Lee**[1], **Ben Ben Liao**[1],
Jiezhong Qiu[1,2], **Qiming Sun**[1], **Jie Tang**[2], **Shengyu Zhang**[1,3]
[1]Quantum Lab, Tencent Co. Ltd.,
[2]Department of Computer Science and Technology, Tsinghua University,
[3]Department of Computer Science and Engineering, CUHK.
{gycchen,kimhsieh,cheekonglee,bliao,qssun,shengyzhang}@tencent.com
qiujz16@mails.tsinghua.edu.cn,jietang@tsinghua.edu.cn

Abstract

We introduce a new molecular dataset, named AlChemy, for developing machine learning models useful in chemistry and materials science. The dataset, as of the workshop submission date, comprises of 12 quantum mechanical properties of 67,000 organic molecules up to 12 heavy atoms, sampled from the GDB database. The Alchemy dataset expands the volume and diversity of existing molecular datasets. Our preliminary benchmark of the state-of-the-art machine learning models on AlChemy clearly manifests the usefulness of new data in validating and developing machine learning models.

Recent advances in machine learning (ML) techniques have proven immensely useful for a broad range of applications including natural language processing (Shen et al., 2017), computer vision (He et al.) and strategic plannings (Silver et al., 2016; 2017) etc. These remarkably successful demonstrations have drawn high interests from the physical and biological science communities. For instance, ML-enhanced abilities to predict molecular properties with high precision (Gilmer et al., 2017; Liao et al., 2019), efficient generations of novel molecular structures (Jin et al., 2018), and better strategies in synthesis planning (Segler et al., 2018) and rectrosynthesis (Segler et al., 2017) will significantly accelerate drug design and novel material discovery (Sanchez-Lengeling & Aspuru-Guzik, 2018; Gómez-Bombarelli et al., 2018). However, the scarcity of high-quality datasets and the lack of transferability of ML techniques impede the wide-scale adoption of molecular ML techniques in practice.

The importance of supervised information for ML development cannot be understated. The ImageNet (Deng et al., 2009), a collection of more than 1.5 million labelled images distributed over $1,000$ classes, facilitates the development of new model such as ResNet (He et al.) that surpasses human performance in image recognition. In another instance, the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), a reading comprehension dataset consisting of $150,000$ questions and answers, is critical to the development of a powerful language representation model BERT (Devlin et al., 2019). Recognizing the importance of abundance of datasets, the chemistry community has recently compiled a comprehensive collection of benchmarking datasets, the MoleculeNet (Wu et al., 2018), for supervised learning tasks. Despite this effort by the Pande group at Stanford, the amount of data in the MoleculeNet is often inadequate in comparison to the typical size of ML training datasets. For instance, there are less than 150K molecular entries for training models to predict the quantum mechanical properties of small organic molecules. The biggest dataset QM9 within MoleculeNet is further restricted to curating molecules composed of Hydrogen (H), Carbon (C), Nitrogen (N), Oxygen (O) and Florine (F). A better data variety, such as presence of more atom types, can help to more thoroughly investigate and improve some aspects of ML models such as transferability and generalibility.

Herein we report a dataset that contains the quantum mechanical properties of around 67,000 organic molecules with up to 12 heavy atoms (C, N, O, S, and halogens including F) from the GDB-17 MedChem database (Ruddigkeit et al., 2012). We name our collection the AlChemy dataset.

[*] Authors are ordered alphabetically

Table 1: Dataset Details: number of molecules and tasks

Dataset	Data Type	#Tasks	#Molecules	Rec-Split	Maximum #Heavy Atoms
QM7	SMILES, 3D coordinates	1	7,160	Stratified	7
QM7b	3D coordinates	14	7,210	Random	7
QM8	SMILES, 3D coordinates	12	21,786	Random	8
QM9	SMILES, 3D coordinates	12	133,885	Random	9
AlChemy	SMILES, 3D coordinates	12	**67,000**	Random	**12**

The quantum mechanical properties are calculated with the Python-based Simulations of Chemistry Framework (PySCF) (Sun et al., 2018). As compared to the full GDB-17 database, the MedChem subset contains molecules that are more suitable for medicinal chemistry, based on functional group and complexity considerations. Therefore, the AlChemy dataset is significantly more comprehensive than existing quantum chemistry datasets, and the molecules are considered more drug-like. AlChemy dataset could serve for evaluation, benchmarking and development of ML methods for applications in chemistry and materials science.

1 Related Work

There are only a few molecular datasets specifically built for benchmarking ML models for chemistry and material applications. Recently, a collection of various datasets have been compiled and collectively branded as the MoleculeNet. These datasets are furthered divided into different categories: quantum mechanics, physical chemistry, biophysics and physiology. In this work, we report our contributions of new datasets for multi-task learning of quantum mechanical properties of organic molecules. Hence, only the subset of quantum mechanical databsets within the MoleculeNet is summarized below.

QM7/QM7b The QM7 dataset contains 7,165 molecules, a subset of GDB-13, with at most 7 C,N,O,S atoms. The learning task is to predict a single electronic property, the atomization energy given the input feature in Table 1. All the physical properties listed in the dataset were calculated with the ab-initio density function theory at the level of PEB0/tier2 basis set. QM7b is an extension. 13 additional properties are computed at different levels of accuracy (ZINDO, SCS, PBE0, GW). The data is further expanded to 7,211 molecules.

QM8 This dataset provides electronic properties of 21,800 molecules comprising up to 8 C,N,O,F atoms, a subset of GDB-17. Not only ground-state electronic properties but also four excited-state properties at different levels of accuracy (TDDFT using PBE0 and CAM-B3LYP and CC2) are computed for multi-task learning.

QM9 This dataset contains geometric, energetic, electronic and thermodynamics (totalling 13) properties for 134,000 organic molecule comprising up to 9 non-hydrogen atom within the GDB-17 database. All physical properties are computed at the accuracy of B3LYP/6-31G(2df,p) based DFT.

2 The AlChemy Dataset

2.1 Molecular Representations: SMILES and 3D Geometry

Intuitively, a molecule is a stable 3D configurations of atoms connected by chemical bonds. The essential features of a molecule can be encoded into a molecular graph with each vertex representing an atom and each edge representing a bond. Simplified Molecular-Input Line-Entry System (SMILES) is a text-based data structure to store molecular graphs on computers. While SMILES and graphs give a minimalist approach to describing molecules, one needs the 3D coordinates of all atoms in a stable configuration for quantum chemistry calculations. This process of recovering 3D geometry from the SMILE string is called the geometry optimization.

Table 2: Calculated properties. Last two entries are not present in the QM9 dataset

Property	Unit	Mean Relative Error	Meam Absolute Error
Rotational constants	GHz	-	-
Dipole moment (mu)	Debye	0.1928	0.6401
Polarizability (alpha)	a_0^3	0.0104	0.7610
HOMO	E_h	0.0244	0.0060
LUMO	E_h	0.2114	0.0061
gap	E_h	0.0275	0.0073
$\langle R^2 \rangle$ (R2)	a_0^2	0.0404	52.3778
Zero point energy (zpve)	E_h	0.0021	0.0003
Internal energy (U0)	E_h	0.0006	0.2486
Internal energy at 298.15 K (U)	E_h	0.0006	0.2486
Enthalpy at 298.15 K (H)	E_h	0.0006	0.2486
Free energy at 298.15 K (G)	E_h	0.0006	0.2486
Heat capacity at 298.15 K (Cv)	$E_\mathrm{h}\mathrm{K}^{-1}$	0.0038	3.8×10^{-7}
Dipole moment vector	-	-	-
3×3 Polarizability tensor	-	-	-

2.2 Detail for the Data Generation

The molecular properties were calculated using PySCF's implementation of the DFT Kohn-Sham method at the B3LYP level with the basis set 6-31G(2df,p). The quantum chemistry model B3LYP/6-31G(2df,p) was validated in an early work (Ramakrishnan et al., 2014) for small organic molecules. Building a complementary dataset along the line of QM-series introduced in Sec. 1, we computed the following categories of molecular properties: ground state equilibrium geometry, ground state electronic properties, and ground state thermochemical properties. The detail is summarized in Table 2.

In the remaining sections, we clarify a few technical aspects of our workflow for interested experts in quantum chemistry. The equilibrium geometry was optimized in three passes. We first used OpenBabel (O'Boyle et al., 2011) to parse SMILES string and built the Cartesian coordinates with MMFF94 force field optimization. The second pass employed the HF/STO-3G theory (incorporate quantum mechanical effects) to generate a preliminary geometry. In the final pass of geometry relaxation, we used the B3LYP/6-31G(2df,p) model with the density fitting approximation for electron repulsion integrals. The auxiliary basis cc-pVDZ-jkfit (Weigend, 2002) is employed in density fitting to build the Coulomb matrix and the HF exchange matrix.

When computing the ground state electronic properties and ground state thermochemical properties, the density fitting technique was disabled in our dataset generation program. In addition to the molecular properties proposed by the original QM9 dataset, we also report the dipole moment vector, the 3×3 polarizability tensor and the atomic charges obtained by the meta-Lowdin population analysis (Sun & Chan, 2014). The advantage of meta-lowdin population analysis, like other advanced population analysis prescriptions (Knizia, 2013), is that the obtained atomic charges possess good transferability in different molecular environments.

2.3 Comparison to the QM9 dataset

In AlChemy dataset, we reproduced results for a random selection of 28,732 molecules from QM9 and calculated properties for another selection of 38,268 molecules up to 12 heavy atoms in the GDB-17 MedChem. The running time statistics and an example of geometry optimization can be found in Appendix. The reproduction of selected QM9 data helps to validate the accuracy of our computations. We draw attention to the third and fourth columns of Table 2. The mean relative error and mean absolute error refers to the disagreement between our results and the original QM9. Our calculations agree well with the original QM9 given the fact that different definitions of B3LYP functionals were used in these calculations. Nevertheless, we notice a few disparities. The relative error on LUMO energy is less a concern as we get a good agreement on the HOMO-LUMO energy gap. It is well-known that the two definitions of B3LYP functionals tend to systematically shift orbital energies in these calculations, and the energy gap should be a more appropriate comparison criterion. However, inconsistency on mu and R2 constitutes more serious indications of disagree-

Table 3: Performance Comparison (MAE)

	QM9			AlChemy
	MPNN[2] (Gilmer et al., 2017)	MPNN (Wu et al., 2018)	MPNN (our impl.)	MPNN (our impl.)
mu	0.03	-	0.5316	2.7448
alpha	0.092	-	0.4171	17.9973
HOMO	0.04257	-	**0.0033**	0.0173
LUMO	0.03741	-	**0.0034**	0.0261
gap	0.0688	-	**0.0048**	0.0262
R2	0.18	-	25.4747	362.8817
ZPVE	0.001524	-	**0.0003**	0.0186
U0	0.01935	-	0.1200	98.3374
U	0.01935	-	0.1201	98.5954
H	0.01677	-	0.0800	92.4926
G	0.01892	-	0.2802	98.2963
Cv	0.04	-	0.1422	33.6612
Average	0.0472245	∼2.7[3]	2.2648	67.0912

ments. We unambiguously trace this inconsistency back to the complexity of searching the optimal molecular conformation among enormous molecular geometry local minimum. The geometry optimization could be stuck at the sub-optimal molecular conformation in either our dataset or QM9. The large deviation between our dataset and QM9's indicates that R2, as a direct output of molecular geometry, is less predictable. This is also observed in our model training practice (Table 3).

3 The Benchmarking Results

We setup baselines using the Message Passing Neural Network (MPNN) model (Gilmer et al., 2017). Our MPNN implementation is based on the authors' prefer model trained individually on each target. As for hyper-parameters not declared in MPNN paper (Gilmer et al., 2017), we refer to their git repository[1] for default parameter setting. For detailed hyper-parameter setting, we list them in Appendix (Table 4). The code of our implementation will be publicly available after the double-blind review period ends.

In the preliminary part of our experiment, we first reproduce the QM9 experiments for 12 targets listed in (Gilmer et al., 2017) with the MPNN model. It is the state-of-the-art model for all regression tasks of the QM9 dataset. QM9 contains 133,885 molecules. We use 10,000 for validation, 10,000 for testing, and the rest for training.

The primary part of our experiment focuses on testing the transferability of existing ML models in quantum chemistry tasks. Models shall be first trained on QM9, and then tested against a subset of our AlChemy dataset with the same atom types (C, N, H, O, F). This fraction contains 37,178 molecules with exactly 10 heavy atoms. The task of generalization to molecular graphs with more heavy atoms is marked as a difficult problem in (Gilmer et al., 2017).

As for comparison metric, Mean absolute error (MAE) is used as evaluation metric in both experiments. Results are reported in Table 3. For reader's convenience, we list the corresponding results reported in (Gilmer et al., 2017; Wu et al., 2018) as well.

We first focus on the comparison among our MPNN implementation and MPNN in (Gilmer et al., 2017) and (Wu et al., 2018). Our MPNN implementation achieves similar performance to (Wu et al., 2018), but differs from (Gilmer et al., 2017)'s results by two orders of magnitude. We also investigate the generalizability of MPNN to larger molecules (see fifth column in Table. 3). Interestingly, we observe large deviations among all targets — the test MAEs on AlChemy dataset differs from those on QM9 dataset by up to orders of magnitude, which again reveals the difficulty in transferring

[1] https://github.com/brain-research/mpnn

[2] The MAEs of Gilmer et al. (2017)'s MPNN can be calculated as (Error Ratio (from their Table. 2)) × (Chemical Accuracy (from their Table. 5)). Interestd readers can refer to a detailed discussion in Appendix.

[3] Refer to Figure 14 in Wu et al. (2018)

knowledge learned from small molecules to larger molecules, although we only increase the number of heavy atoms by one.

4 Conclusions and Future Work

The process of building the AlChemy dataset has been a rewarding experience. Along this journey, we face the difficulty to optimize molecular geometries efficiently and accurately. We eventually use a combination of two geometry optimization tools to overcome this problem. All the discrepancies between our data and QM9's can be precisely attributed to the disagreements on the optimized geometries. Our preliminary benchmark results clearly manifest the usefulness of the AlChemy dataset. The public availability of this dataset shall inspire further improvement of existing ML models as well as the proposals of entirely new ones.

References

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR '09, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT '19, 2019.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural Message Passing for Quantum Chemistry. In ICML '17, pp. 1263–1272, 2017.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-driven Continuous Representation of Molecules. ACS Central Science, 4(2):268–276, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction Tree Variational Autoencoder for Molecular Graph Generation. ICML '18, 2018.

Gerald Knizia. Intrinsic Atomic Orbitals: An Unbiased Bridge between Quantum Theory and Chemical Concepts. Journal of Chemical Theory and Computation, 9(11):4834–4843, 2013.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated Graph Sequence Neural Networks. ICLR' 16, 2016.

Renjie Liao, Zhizhen Zhao, Raquel Urtasun, and Richard S Zemel. LanczosNet: Multi-Scale Deep Graph Convolutional Networks. In ICLR '19, 2019.

Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open babel: An open chemical toolbox. Journal of Cheminformatics, 3(1):33, 2011.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In EMNLP '16, 2016.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. Scientific Data, 1:140022, 2014.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. Journal of Chemical Information and Modeling, 52(11):2864–2875, 2012.

Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. Science, 361(6400):360–365, 2018.

Marwin Segler, Mike Preuß, and Mark P Waller. Towards "Alphachem": Chemical Synthesis Planning with Tree Search and Deep Neural Network Policies. In ICLR '17 Workshop, 2017.

Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. Nature, 555(7698):604, 2018.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. Reasonet: Learning to Stop Reading in Machine Comprehension. In KDD '17, pp. 1047–1055. ACM, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. Nature, 529(7587):484, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the Game of Go without Human Knowledge. Nature, 550(7676):354, 2017.

Qiming Sun and Garnet Kin-Lic Chan. Exact and optimal quantum mechanics/molecular mechanics boundaries. Journal of Chemical Theory and Computation, 10(9):3784–3790, 2014.

Qiming Sun, Timothy C Berkelbach, Nick S Blunt, George H Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D McClain, Elvira R Sayfutyarova, Sandeep Sharma, et al. Pyscf: the python-based simulations of chemistry framework. Wiley Interdisciplinary Reviews: Computational Molecular Science, 8(1):e1340, 2018.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. ICLR '15, 2015.

Florian Weigend. A Fully Direct RI-HF Algorithm: Implementation, Optimised Auxiliary Basis Sets, Demonstration of Accuracy and Efficiency. Phys. Chem. Chem. Phys., 4:4285–4291, 2002.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a Benchmark for Molecular Machine Learning. Chemical Science, 9(2):513–530, 2018.

5 Appendix

5.1 Hyper-parameters

We list detailed hyper-parameter settings of our MPNN implementation in Tabel. 4.

5.2 More Details on AlChemy Data Generation

Figure. 1 shows AlChemy's running time statistics; Figure. 2 shows an example of geometry optimization process from AlChemy dataset.

5.3 Conversion of Units

As clearly specified in Table 2, we adopt the atomic units for all numerics in the AlChemy dataset. This choice is consistent with the QM9 dataset. However, we caution that the MAEs in the second column of Table 3, quoted from Gilmer's work (Gilmer et al., 2017), are given in a different set of units. For instance, one has to divides all energy-related MAEs except the one for the zero-point vibrational energy by 27 to convert the numbers into the atomic units. After appropriate conversions, the disagreements between the actual numbers of Gilmer's original work and our MPNN results will be more pronounced.

Table 4: Hyper-parameters.

Category	**Hyper-parameter**	**Value**
Preprocessing	edge features	bond type + raw distance
	node features	Table 1 in Gilmer et al. (2017)
	use Mulliken partial charges	No
Message Passing Function (Edge Network)	message passing steps T	6
	edge network layers	4
	edge network hidden dim	50
	use multiple towers	No
Update Function (GRU, as in GG-NN (Li et al., 2016))	node hidden units d	50
	GRU layer	1
Readout Function (set2set (Vinyals et al., 2015))	set2set computations M	12
	output NN hidden layer	1
	output NN hidden units	200
Training	initial learning rate	0.00013
	learning rate decay factor	0.5
	optimizer	Adam
	batch size	20
	traing epochs	540

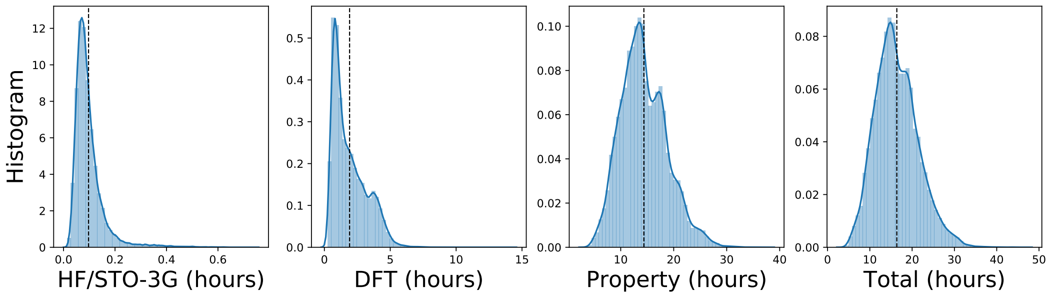
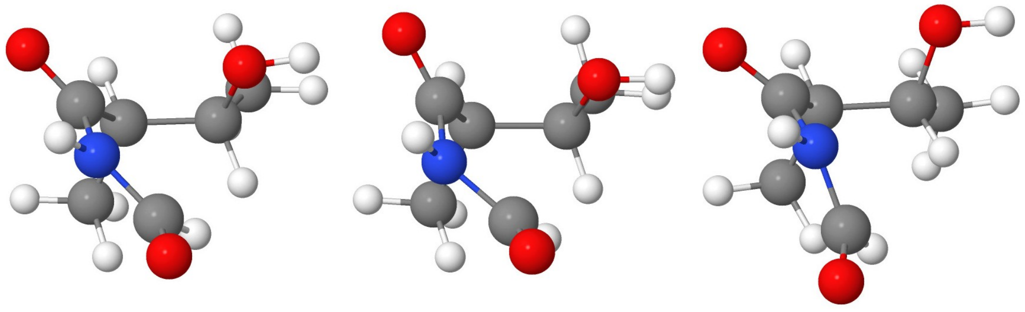


Figure 1: AlChemy's running time statistics for molecules (with 10 heavy atoms) from GDB-17 MedChem. Each step takes 0.1/1.9/14.4 hours on average, respectively. The average total running time is 16.4 hours.



(a) OpenBabel geometry (b) HF/STO-3G geometry (c) DFT geometry

Figure 2: An example for molecule CC(O)C(C)C(=O)NC=O, a random sample from AlChemy dataset.