Geometry Aware Deep Learning for Integrated Closed-shell and Open-shell Systems

Beom Seok Kang [*1] **Vignesh C Bhethanabotla** [*1] **Mohammadamin Tavakoli** [2] **William A Goddard III** [1]
Anima Anandkumar [2]

Abstract

Simulations of chemical systems rely on calculation of their potential energy surfaces (PES), i.e., a function which returns the energy of a system under study. The electronic structure of a molecule may be closed-shell or open-shell, where either all electron spins are paired, or one or more electrons are unpaired in spin, respectively. While the cost of quantum-chemistry calculations can be reduced by assuming a closed-shell electronic structure and removing the necessity of the spin degree of freedom, it is often important to consider systems with unpaired spins, i.e. open-shell, such as in radical chemistry or description of chemical reactions. Here, we propose an extension for OrbNet-Equi, an equivariant deep-learning quantum mechanical approach to representing chemical systems at the electronic structure level. By utilizing a spin-polarized treatment of the underlying semi-empirical quantum mechanics featurization, OrbNet-Equi can describe both closed and open-shell electronic structures. We test the efficacy of this new representation with representative datasets.

1. Introduction

Predicting molecular properties is a highly important task in various areas such as drug discovery and material science. The space of possible molecular structures is huge, and high-throughput methods for exploring these spaces is critical. (Leelananda & Lindert, 2016).

Density functional theory (DFT) is a standard method to approach the many-electron system. Though fairly accurate for many systems of interest, the relatively high cost and cubic scaling with respect to the number of basis functions, which scales with the system size, limit its applicability. Hence, predicting the properties of a large number of molecules using DFT can often be limited due to the severe computational cost. Semi-empirical methods circumvent some of the cost by approximating some of the physical interactions to parameterized functions, but suffer from a reduced accuracy.

In the Born-Oppenheimer approximation, a chemical system is separated into its nuclear degrees of freedom and its electronic degrees of freedom, the former of which is treated classically and the latter of which is treated quantum mechanically. Solving the Schödinger equation for the electrons yields the energy and orbitals of the system as a function of the coordinates of its nuclei. Electrons have both spatial degrees of freedom and a spin degree of freedom, the latter of which is binary up or down. Being fermions, they obey the Pauli exclusion principle, which prevents any two electrons from occupying the same state. Practically, this means that two electrons can occupy the same spatial orbital only if they have opposite spins. In this approximate description, the electronic structure can be closed-shell or open-shell. The closed-shell system assumes that all electron spins are paired, which allows for a more efficient treatment where only spatial degrees of freedom need to be considered in the solution of the Schrödinger equation. The open-shell system is the opposite of this, where there exist unpaired electrons, and necessitates a treatment of both the spatial and spin degrees of freedom in the calcualted orbitals.(Szabo & Ostlund, 1989).

Open-shell systems are especially important in many areas of chemistry, such as describing radicals, reactive intermediates, transition metal complexes, and other important applications which all encounter unpaired electrons. Importantly, from the perspective of a representation for machine-learning on molecular systems, calculating the energy of the open or closed-shell state of a system is not dependent on its atomic geometry, and so a description of the spin state must be incorporated to properly account for this structure.

[*]Equal contribution [1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, USA [2]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, USA. Correspondence to: Beom Seok Kang <bkang@caltech.edu>, Vignesh C Bhethanabotla <vbhethan@caltech.edu>.

Accepted as an extended abstract for the Geometry-grounded Representation Learning and Generative Modeling Workshop at the 41st International Conference on Machine Learning, ICML 2024, Vienna, Austria.

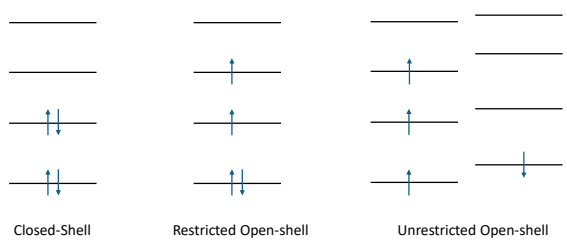

Figure 1. Illustration of energy levels of orbitals and electrons occupying the orbitals. Closed-shell assumes all electrons to be paired, restricted open-shell constrains spatial orbitals to be identical for different spins, and unrestricted open-shell considers the spin polarization, leading different spin orbitals at the same spatial orbital to have different energies.

While extensive research has been done on learning molecular representations and predicting properties for closed-shell molecules, there has been comparatively little focus on open-shell molecules despite their significance. Most machine learning models that predict molecular properties rely on molecular coordinates and element numbers as inputs. This approach limits the generalizability of the models to open-shell molecules, as they cannot distinguish between closed-shell and open-shell states. Incorporating additional embeddings could address this issue, but an effective method for representing molecules with different spin states remains unclear.

OrbNet-Equi is a deep learning model that learns the mapping between molecular electronic structure and physical quantities (Qiao et al., 2022). The previous work from Qiao et al. (2022) generates inputs using the semi-empirical GFN1-xTB (Grimme et al., 2017) method. The model has shown great potential in increasing accuracy to the level that is on par with DFT/B3LYP level of theory. Moreover, its equivariantly constructed neural network allows the accurate prediction of some equivariant properties, such as forces and dipole moments. However, the current featurization of OrbNet-Equi learns on top of is strictly a closed-shell representation of the molecular system.

Here, we present our plans for generalizing the OrbNet-Equi structure to integrate the training for closed-shell and open-shell molecules. We discuss our method to expand this to training open-shell molecules along with closed-shell molecules and how they can be distinguished by the model. We expect that such an attempt can provide a good insight in exploring the chemical space with different compositions and conformations, and expand the capability of this model to exploring more diverse chemical systems.

2. Related Works

2.1. Open-shell System Learning

In this section, we discuss some previous work that addressed the prediction of open-shell systems (some together with closed-shell systems) using machine learning.

Lemm et al. (2021) implemented a machine learning model Graph-To-Structure (G2S) to reconstruct atomic coordinates by predicting the interatomic distances, from bond network and stoichiometry, which allows to bypass of the costly energy minimization task of force-fields or *ab initio* methods. However, this work mainly focuses on structure optimization, and therefore has a different goal to ours and cannot be compared directly.

Cheng et al. (2022a) successfully implemented a machine learning method called molecular-orbital-based machine learning (MOB-ML). This work uses Fock (**F**), exchange (**K**), and Coulomb (**J**) matrices but uses the localized molecular orbitals instead of the atomic orbital basis. The method additionaly uses Gaussian Process Regression (GPR) in contrast to a deep-learning approach. Further, the MOB-ML approach utilizes the Hartree-Fock (HF) method for generating its MO features, which are higher cost than DFT calulations, and usually targets higher level theory labels, such as coupled-cluster or multi-reference methods. Such an approach shows accurate predictions with several different datasets, for both closed-shell and open-shell.

In addition, Unke et al. (2021) introduced SpookyNet, which is a neural network model that constructs force fields via self-attention. SpookyNet requires four inputs: atomic numbers, Cartesian coordinates, total charge, and total angular momentum, i.e. the number of unpaired electrons. The inputs are processed into different representations, and then they are passed through the neural network to predict energy. This approach applies empirical augmentations to extrapolate beyond the training data, showing good performance on different datasets with both closed-shell molecules and open-shell molecules. However, SpookyNet has weaknesses in predicting the properties of unknown molecules and conformations.

2.2. OrbNet

OrbNet is a deep learning framework that was introduced to predict the molecular energy from atomic orbital (AO) features (Qiao et al., 2020). The model employs AO features, represented by quantum chemical matrices produced while self-consistent field (SCF) convergence, as inputs to it. OrbNet uses AO features which are evaluated in a symmetry-adapted atomic-orbital (SAAO) basis. The AO features are then encoded into graph-structured data which then are applied to a model of which the graph neural network (GNN) architecture is adopted. The data represented on the graph

are passed through a neural network, and the output tensor is decoded and summed to yield the final output, which is the molecular energy. More detailed explanations of the theory and structure of the model can be found in (Qiao et al., 2020).

3. Method

3.1. OrbNet-Equi

OrbNet-Equi utilizes AO features and allows prediction of the translation-rotation ($E(3)$) equivariant properties of molecules such as atomic forces and invariant properties such as molecular energy (Qiao et al., 2022).

OrbNet-Equi uses matrices that are generated using the GFN1-xTB method (Grimme et al., 2017). Specifically, the matrices that are calculated using GFN1-xTB are the Fock (**F**), density (**P**), the Hamiltonian (**H**), and the overlap (**S**) matrices which are put into a vector ($\mathbf{T} = [\mathbf{F}, \mathbf{P}, \mathbf{H}, \mathbf{S}]$) as input. The matrices are generated from their corresponding operators and orbitals. For example, the Hamiltonian matrix **H** is formulated by

$$(\mathbf{H})_{AB}^{n,l,m;n',l',m'} = \langle \Phi_A^{n,l,m} | \hat{\mathcal{H}} | \Phi_B^{n',l',m'} \rangle, \tag{1}$$

where A and B indicate different atoms, (n, l, m) and (n', l', m') are the basis sets for the atoms, and $\hat{\mathcal{H}}$ is the Hamiltonian operator.

Hole excitation and particle excitation matrices are also attempted to be used in the original work, but in this work we only use **F**, **P**, **S** and **H** matrices as a starting point. GFN1-xTB is a semiempirical method that allows for a rapid calculation but is less accurate than density functional theory (DFT) methods. OrbNet-Equi applies the "delta-learning" strategy (Ramakrishnan et al., 2015), which is a strategy that learns the discrepancy between the predicted properties calculated at a lower (i.e., more approximate) level of theory and a higher (but more costly) level of theory. Specifically, the original OrbNet implementation tested on the delta labels between the semi-empirical GFN-xTB and a higher level DFT treatment. Since the properties can be rapidly calculated from the semiempirical GFN1-xTB method, the relative inaccuracy can be obtained several orders of magnitude quickly compared to that done purely by DFT methods, which often require a much greater amount of calculations.

Although the original article (Qiao et al., 2022) focuses on the use of the closed-shell GFN1-xTB method to calculate input matrices, the structure of OrbNet-Equi is not limited to it. To study and learn the mapping between the open-shell system and the molecular properties, we apply OrbNet-Equi architecture but with different data representations and outputs. The model we train learns the mapping $\mathcal{F}$ between the open-shell system's matrices vector $\mathbf{T} = [\mathbf{F}, \mathbf{P}, \mathbf{H}, \mathbf{S}]$ and the label molecular property **y** that is either generated by simulation or estimated by experiments, i.e. the goal is to perform,

$$\min_{\mathcal{F}} \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}), \tag{2}$$

where $\mathcal{L}$ is the loss function for training and $\hat{\mathbf{y}} = \mathcal{F}(\mathbf{T})$ is the prediction of the label molecular property **y** from the model.

We also highlight that OrbNet-Equi is capable of learning equivariant properties, such as forces, which are defined by

$$\vec{F}_u = -\nabla_{\vec{r}_u} E, \tag{3}$$

where $\vec{F}_u$, $\vec{r}_u$, E are the force vector for atom u, the position vector of atom u, and the energy of the molecule, respectively. Since force is equivariant to translation and rotation, to predict it accurately and correctly, the neural network must also be equivariant to them:

$$\mathcal{R} \cdot \mathcal{F}(\mathbf{T}) = \mathcal{F}(\mathcal{R} \cdot \mathbf{T}), \tag{4}$$

where $\mathcal{R}$ is an arbitrary rototranslational operation. We note that invariance is a special case of equivariance where the output does not change with rotation of the input. Such rotation on matrices results in a predictable change in their elements. To elaborate, for each block that contains interaction between atom A's orbital with angular momentum quantum number l and atom B's orbital with angular momentum quantum number l', the transformation $\mathcal{R} \cdot \mathbf{T}$ results as follows,

$$(\mathcal{R} \cdot \mathbf{T})_{AB}^{l;l'} = \mathcal{D}^l(\mathcal{R})(\mathbf{T})_{AB}^{l;l'} \mathcal{D}^{l'}(\mathcal{R})^\dagger, \tag{5}$$

where $\mathcal{D}^l(\mathcal{R})$ is the Wigner-D matrix of degree l, for the given operation $\mathcal{R}$. The dagger symbol represents the Hermitian conjugate. The design of OrbNet-Equi satisfies the equivariance by successfully implementing an equivariant neural network, namely $\mathcal{F}$. Detailed explanations of the architecture of equivariant neural networks are addressed in the article (Qiao et al., 2022). A brief illustration of the overview is shown in figure 2.

3.2. Generating atomic orbital features for integrated closed-shell and open-shell systems

As open-shell systems take into account the unpaired electrons, the different spin orbitals are also considered. Hence, each spatial orbital can be split into two spin orbitals. Calculations in the spin basis can be converged either in a restricted or unrestricted treatment. In the former, the overall spin state is specified by constraining the spatial orbitals

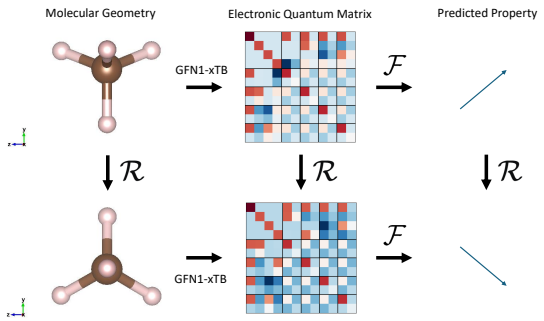

Figure 2. Illustration of the equivariance in OrbNet-Equi. The geometry of the methane molecule is rotated by 90° in a counter-clockwise direction about the z-axis, which is denoted here by $\mathcal{R}$. The example matrices are the Fock matrices ($\mathbf{F}$) generated by the GFN1-xTB method using each geometry. Black borders show the grouping of the elements to their atomic centers. The matrices are then forwarded to the equivariant neural network $\mathcal{F}$ to get the predicted equivariant properties. The resulting equivariant property may include vector quantities such as forces and dipole moment and invariant scalar quantities such as energy.

to be either singly or doubly occupied. In the latter, all orbitals are allowed to relax the spin degree of freedom, and the spin state is specified by the initial guess state plugged into the self consistent field method. In this work, we use the implementation of the unrestricted SCF procedure for a spin-polarized calculation in the GFN1-xTB form, referred to as the spGFN1-xTB method (Neugebauer et al., 2023).

The current OrbNet-Equi architecture utilizes the AO features generated directly from the closed-shell GFN1-xTB method calulated in the restricted closed-shell configuration. This means that all electrons are paired, which allows for the simplification of the orbital basis to only contain the spatial degrees of freedom, and assumes that all occupied orbitals contain 2 electrons, one for each spin. In contrast, an open-shell calculation will allow for unpaired spins. For different spins, we shall formulate different Roothaan equations, i.e.,

$$\mathbf{F}^{\alpha}\mathbf{C}^{\alpha} = \mathbf{S}\mathbf{C}^{\alpha}\epsilon^{\alpha}, \tag{6}$$

$$\mathbf{F}^{\beta}\mathbf{C}^{\beta} = \mathbf{S}\mathbf{C}^{\beta}\epsilon^{\beta}, \tag{7}$$

where the superscripts α and β are the different spins, $\mathbf{C}$ is the orbital coefficients matrix, and ϵ is the diagonal matrix of orbital energies. Hence, for the unrestricted open-shell system, instead of a single Fock matrix generated from spatial orbitals interactions, we get two Fock matrices for different spins (Szabo & Ostlund, 1989). Similarly, we get two density matrices for different spins, which each element is given by,

$$P^{\alpha}_{\mu\nu} = \sum_{a}^{N^{\alpha}} C^{\alpha}_{\mu a}(C^{\alpha}_{\nu a})^{*} \tag{8}$$

$$P^{\beta}_{\mu\nu} = \sum_{a}^{N^{\beta}} C^{\beta}_{\mu a}(C^{\beta}_{\nu a})^{*} \tag{9}$$

where P and C are the individual elements of the density matrix and the orbital coefficients matrix, respectively, the subscripts represent the position of elements, and N^{α} and N^{β} are number of electrons with α spin and β spin, respectively (Szabo & Ostlund, 1989).

This generally applies to closed-shell systems as well. However, for closed-shell systems, matrices are identical for different spins. More precisely, we get $\mathbf{F}^{\alpha} = \mathbf{F}^{\beta} = \mathbf{F}$, and $\mathbf{P}^{\alpha} = \mathbf{P}^{\beta} = \frac{1}{2}\mathbf{P}$ for closed-shell molecules. In this way, we can create a consistent representation for both closed-shell and open-shell systems. With this representation, now our vector of atomic orbital features is $\mathbf{T} = [\mathbf{F}^{\alpha}, \mathbf{F}^{\beta}, \mathbf{P}^{\alpha}, \mathbf{P}^{\beta}, \mathbf{S}, \mathbf{H}]$ All AO features can be generated using the spGFN1-xTB method.

Another possibility is to use open-shell extended matrices, where then each row and column represent spin atomic orbitals instead of spatial ortibals. Since for each spatial orbital it can split into two spin orbitals, the extended matrices will have twice larger numbers of rows and columns. Using the larger matrices, we can continue using the same structure of the model. However, a possible problem that this method might entail is that this method is expected to be more expensive than the above method for separating matrices with different spins. Also, there will exist many 0's in every matrices, which may cause inefficiencies.

Both ways allow closed-shell molecules and open-shell molecules to be represented in a consistent system. Moreover, they contain the same amount of information, and are not limited to be used for GFN1-xTB matrices. The examples of both representations are shown in figure 3. We plan to attempt both methods. For the closed-shell system, the AO features are generated using the xtb implementation in the tblite package (Ehlert, 2024). For the open-shell calculations, the spGFN1-xTB method, implemented in the same package, will be used (Neugebauer et al., 2023).

3.3. Datasets

For integration, we first plan to train the model on energy predictions with both an open-shell system dataset, QM-Spin (Schwilk, Max et al., 2020; Cheng et al., 2022b), and a closed-shell system dataset, QM9 (Ramakrishnan et al., 2014).

The QM9 data set consists of 134,000 stable small organic molecules with H, C, N, O, and F atoms (Ramakrishnan

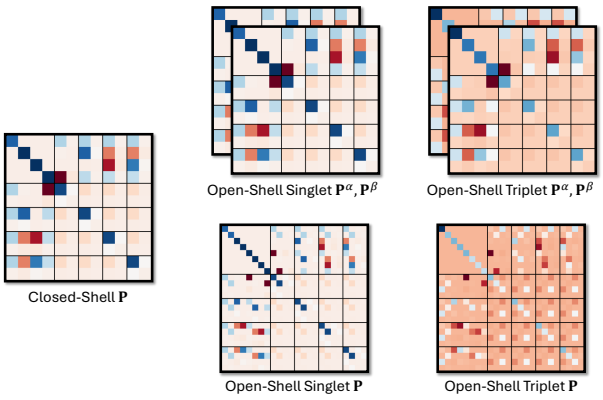

Figure 3. Density matrices (**P**) generated using GFN1-xTB for closed-shell methane (CH_4) molecule (left), for open-shell singlet methane molecule with separate matrices for different spins (upper middle) and a single matrix for both spins (lower middle), and for open-shell triplet methane molecule with separate matrices for different spins (upper right) and a single matrix for both spins (lower right).

et al., 2014). For each molecule, properties that are calculated using DFT with B3LYP/6-31G(2df,p) level of theory are provided, which include: dipole moment, HOMO energy, LUMO energy, internal energies at 0 K and 298.15 K, and others. Specifically, we focus on the training model with internal energy at 0 K.

The QMSpin dataset consists of 4,980 singlet and 7,834 triplet state carbene molecules' data generated from the QM9 dataset, which each data contains: atomic numbers, molecular geometry (coordinates of each atom), singlet energy, and triplet energy (Schwilk, Max et al., 2020; Cheng et al., 2022b). The geometries are optimized via restricted open-shell B3LYP/def2-TZVP, and most of the energies are obtained at MRCISD+Q-F12/cc-pVDZ-F12 (Cheng et al., 2022a).

However, since QM9 and QMSpin use the different levels of theories to calculate energies, they cannot be trained together. Hence, a reasonable approach would be, to train a model with QM9, i.e. closed-shell molecules, so that it can learn the part of underlying physics from the abundant closed-shell system data, and to fine-tune the model with open-shell systems such as QMSpin to learn the remaining part. However, after fine-tuning, the model will only be valid to predict QMSpin molecules. This may not be universally applicable as QMSpin consists of carbene molecules, and hence it is probable that the trained model can optimize on carbene structures only. Therefore, there arises the need of general open-shell molecules dataset which is not limited to certain structures. To our best knowledge, there does not exist a sufficiently large dataset for such need.

We also plan to train our model for the prediction of forces, which is also an essential task for creating a potential energy surface with respect to the atomic coordinates. There exist several datasets for this that provide both energies and forces, including rMD17 (Christensen, Anders & von Lilienfeld, O. Anatole, 2020) and SPICE (Eastman et al., 2023) which are datasets with closed-shell molecules, and AIMNet-NSE (Zubatyuk et al., 2021) with open-shell molecules.

3.4. Training

For training, we plan to use the same model architecture and hyperparameters as those used for the previous OrbNet-Equi (Qiao et al., 2022). In addition, we plan to use the Adam optimizer (Kingma & Ba, 2014) with a linear warm-up with 100 epochs followed by cosine annealing with 200 epochs with a maximum learning rate of 5×10^{-4}. We plan to use a batch size of 64 for training. Moreover, the loss function $\mathcal{L}$ we plan to use for the integrated learning is smoothL1Loss, which is a loss function with a quadratic slope below a certain error and a linear slope above the error.

The ratio of open-shell and closed-shell molecules in the training dataset is to be determined in such a way that the model can capture the physics for both closed-shell and open-shell systems well. The training can be done simultaneously using different datasets of those which share the same level of theory and the basis set, or it can be done firstly with closed-shell molecules, of which the amount of data is greatly abundant, then the resultant model can be fine-tuned with open-shell molecules to improve accuracy for predicting open-shell systems, which is applicable to currently available datasets.

4. Conclusion and Future Works

We presented our plan of generalizing the existing OrbNet-Equi framework for integrating the training for open-shell and closed-shell systems. We introduced the two possible ways to represent closed-shell and open-shell systems. They both are consistent and generalizable for both systems, and are able to distinguish between molecules with different total spins. With the given framework and the ways of representations, we plan to apply several changes to OrbNet-Equi to train on the closed-shell and open-shell systems.

We plan to train the model for predicting the energies and forces of different molecules for different systems. We propose two possible strategies; one is simultaneous training and the other is closed-shell system training followed by fine-tuning with open-shell system data. We summarize the available datasets that can be used.

References

Cheng, L., Sun, J., Deustua, J. E., Bhethanabotla, V. C., and Miller, T. F. Molecular-orbital-based machine learning for open-shell and multi-reference systems with kernel addition Gaussian process regression. *The Journal of Chemical Physics*, 157(15):154105, October 2022a. ISSN 0021-9606, 1089-7690. doi: 10.1063/5.0110886.

Cheng, L., Sun, J., Deustua, J. E., Bhethanabotla, V. C., and Miller III, T. F. Criegee, h10 chain, small radicals, water bond dissociation, and qmspin energy datasets with mob features for mob-ml(ka-gpr), 2022b. URL https://data.caltech.edu/records/20200.

Christensen, Anders and von Lilienfeld, O. Anatole. Revised md17 dataset, 2020. URL https://archive.materialscloud.org/record/2020.82.

Eastman, P., Behara, P. K., Dotson, D. L., Galvelis, R., Herr, J. E., Horton, J. T., Mao, Y., Chodera, J. D., Pritchard, B. P., Wang, Y., De Fabritiis, G., and Markland, T. E. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Scientific Data*, 10(1):11, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01882-6. URL https://www.nature.com/articles/s41597-022-01882-6.

Ehlert, S. awvwgk/tblite: Light-weight tight-binding framework. https://github.com/awvwgk/tblite, 2024.

Grimme, S., Bannwarth, C., and 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). *Journal of Chemical Theory and Computation*, 13(5):1989–2009, 2017. doi: 10.1021/acs.jctc.7b00118. URL https://doi.org/10.1021/acs.jctc.7b00118. PMID: 28418654.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Leelananda, S. P. and Lindert, S. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry*, 12:2694–2718, December 2016. ISSN 1860-5397. doi: 10.3762/bjoc.12.267. URL https://www.beilstein-journals.org/bjoc/articles/12/267.

Lemm, D., von Rudorff, G. F., and von Lilienfeld, O. A. Machine learning based energy-free structure predictions of molecules (closed and open-shell), transition states, and solids. *Nature Communications*, 12(1):4468, July 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24525-7. URL http://arxiv.org/abs/2102.02806. arXiv:2102.02806 [physics].

Neugebauer, H., Bädorf, B., Ehlert, S., Hansen, A., and Grimme, S. High-throughput screening of spin states for transition metal complexes with spin-polarized extended tight-binding methods. *Journal of Computational Chemistry*, 44(27):2120–2129, July 2023. ISSN 1096-987X. doi: 10.1002/jcc.27185. URL http://dx.doi.org/10.1002/jcc.27185.

Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R., and Miller, T. F. OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics*, 153(12):124111, September 2020. ISSN 0021-9606, 1089-7690. doi: 10.1063/5.0021955.

Qiao, Z., Christensen, A. S., Welborn, M., Manby, F. R., Anandkumar, A., and Miller, T. F. Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proceedings of the National Academy of Sciences*, 119(31):e2205221119, August 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2205221119.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, August 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Big data meets quantum chemistry approximations: The -machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099. URL https://doi.org/10.1021/acs.jctc.5b00099. PMID: 26574412.

Schwilk, Max, Tahchieva, Diana N., and von Lilienfeld, O. Anatole. The qmspin data set: Several thousand carbene singlet and triplet state structures and vertical spin gaps computed at mrcisd+q-f12/cc-pvdz-f12 level of theory, 2020. URL https://archive.materialscloud.org/record/2020.0051/v1.

Szabo, A. and Ostlund, N. S. *Modern Quantum Chemistry - Introduction to Advanced Electronic Structure Theory*. Dover Publications, 1989.

Unke, O. T., Chmiela, S., Gastegger, M., Schütt, K. T., Sauceda, H. E., and Müller, K.-R. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature Communications*, 12(1):7273, December 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-27504-0. URL https://www.nature.com/articles/s41467-021-27504-0.

Zubatyuk, R., Smith, J. S., Nebgen, B. T., Tretiak, S., and Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *Nature Communications*, 12 (1):4870, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24904-0. URL https://www.nature.com/articles/s41467-021-24904-0.