

## HemeSchNet

**HemeSchNet: A continuous Filter Convolutional Neural Network for prediction of heme cofactor orbital energies**

*J. W. Jones, S. Leipnitz, M. A. Mroginski, Institute for Biomolecular Modeling, Technische Universität Berlin*

### In Short

- Create HemeSchNet Models
- Optimize Hyperparameters
- Predict Heme Orbital Energies

### Previous Project

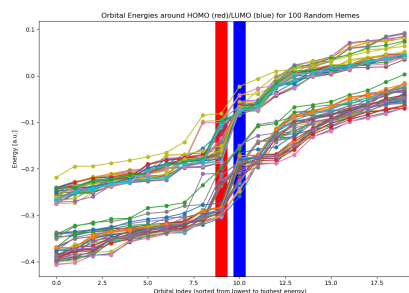
In the previous project titled "Calculation of Crystal Structure Heme Properties for Reaction Prediction", quantum mechanical computations were performed on heme crystal structures and their axial ligands within individual heme protein structures. As described in the original application, these calculations were executed using the Gaussian16 software package with GPU acceleration. Approximately 1,600 structures were intended to be analyzed, largely preserving the in vivo configurations derived from crystal structures, across four distinct states differing in spin and multiplicity. A subsequent proposal for future work was proposed in the first projects application, contingent upon the successful convergence of a sufficient number of calculations. Data was obtained for around 1500 structures, others did default due to initial convergence issues. From these calculations, a database was created.

A preliminary analysis of the data generated from the computations outlined in the original proposal reveals the orbital energies of the ten highest occupied molecular orbitals (HOMO) and the ten lowest unoccupied molecular orbitals (LUMO) for 100 randomly selected heme cofactors, as depicted in Figure 1.

The data was extracted from the Gaussian log-files of the calculations done in the previous project, showing a small part of the results.

### Current Project

HemeSchNet, the model proposed to be trained with the current follow-up project proposal is supposed to be trained on those values. Two clearly distinguishable populations of Heme orbital energies are observable, with one of the populations having significantly higher energies for the molecular orbital



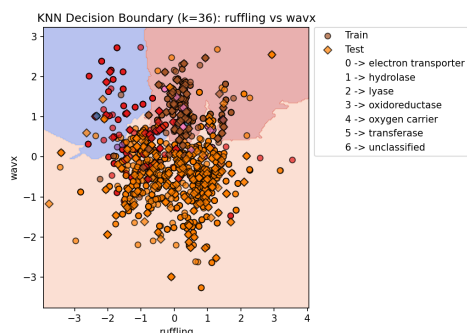
**Figure 1:** Orbital energies for 100 randomly selected heme cofactors were extracted from the Gaussian16 log files. The energies of the orbitals are organized in ascending order, with HOMOs highlighted by red boxes and LUMOs highlighted by blue boxes.

energies compared to the other population. Based on the theory stated in more detail in the original proposal in section 1, the expectation is that these two populations of hemes might also facilitate different types of reactions. Additionally, some heme structures exhibit a different, equally unexpected behavior, where the highest difference in energy between the displayed molecular orbital energies is not between the HOMO and the LUMO, as is also the case with most of the displayed structures, but between the highest occupied molecular orbital and the second highest occupied molecular orbital, or the second highest occupied molecular orbital with the third highest molecular orbital. This could also indicate a particularly interesting behavior for the heme structures in the facilitated reactions.

Initial investigation of the data together with the data obtained from the pyDISH database [?] that describes the calculated heme distortions with the k-Nearest Neighbor method and the best model obtained with cross validation is shown in figure 2, detailing a latent space classification separation.

Originally it was outlined that a comprehensive database of various single-heme-protein heme moiety crystal structures and their corresponding physical parameters would be established. This database is intended to underpin the development of an advanced machine learning model, called HemeSchNet.

HemeSchNet is constructed upon the SchNet architecture [?], which is a type of continuous filter convolutional neural network (CFCNN). CFCNNs are designed to recognize spatial information with arbitrary positioning, representing a significant improvement over traditional convolutional neural networks (CNNs) that are limited to discrete, equidistant



**Figure 2:** Latent space separation of the  $k=36$  neighbors best model from the test evaluation with cross validation on the dataset of distortions and orbital energies around the HOMO and LUMO energies trying to classify the reaction type facilitated by the heme.

grid points with fixed resolutions. Traditional CNNs typically require an increase in model parameters as the number of grid points expands. In contrast, SchNet and other CFCNNs utilize spatial transformations learned by the model to represent arbitrary positions through relative coordinates, thereby mitigating the growth in parameters.

This project aims to exploit the unique capabilities of CFCNNs, particularly the SchNet architecture's status as a physically informed neural network. CFCNNs model energies and forces by leveraging rotational symmetry, which reduces the number of parameters necessary to describe a molecular system at a quantum mechanics-like level. This efficiency is achieved not only through the aforementioned spatial transformations but also via interaction blocks. These interaction blocks model atomic interactions as potentials, becoming increasingly refined with each additional sequential block.

Consequently, the CFCNN architecture offers superior scalability compared to traditional CNNs in two key aspects: first, it achieves higher accuracy by maintaining the inferred transformations of spaces with arbitrary positions throughout all evaluations; second, its physically informed properties, based on inherent symmetries and interaction blocks, enable the model to make more accurate predictions with fewer parameters. Within the machine learning community, various heuristics exist regarding the optimal number of data points relative to the number of model parameters to prevent overfitting and ensure effective learning of the underlying ground truth. However, machine learning models in chemistry often fall short of these heuristics due to data sparsity. This project addresses this challenge through the more efficient utilization of parameters inherent to the CFCNN architecture.

The data generated from the original project will serve as the only publicly available database of phys-

ical parameters for heme moieties derived from crystal structures. Current literature suggests that heme distortions encode information about the heme's environment and its integration within the overall protein structure, ultimately modulating the heme's redox potential. This redox potential is believed to directly influence the type of reactions facilitated by the heme protein. While preliminary investigations indicate that redox potential alone does not solely determine the type of reaction facilitated by the heme moieties, the most significant output of the DFT calculations conducted in the prior project are the orbital energies. These energies provide a more nuanced understanding of how heme structures are modulated at a subatomic level and elucidate the interactions that may be involved in the active center. This detailed information enables a systematic approach to predicting the type of reaction a heme moiety facilitates based solely on its structure.

HemeSchNet is designed to fulfill this objective by predicting physical parameters and potentially inferring the reaction type for an unknown heme structure using only structural input, such as atomic positions and atomic types derived from the number of protons in each atomic core. The preliminary analysis presented demonstrates, that while applying machine learning algorithms to this problem is promising, models that are less computationally intensive lack the capacity to learn sufficient ground truth to accurately predict previously unseen data.

## WWW

<https://www.tu.berlin/biomodeling/ueberuns/leitung>

## More Information

- [1] KT Schütt, HE Saucedo, P-J Kindermans, A Tkatchenko, K-R Müller, *J. Chem. Phys.* **148**, 241722 (2018)
- [2] HX Kondo, Y Kanematsu, G Masumoto, Y Takano, *Database* **2023**, baaa066 (2023)

## DFG Subject Area

201-02