Machine Learning the Fock Matrix in the Atomic Orbital Basis for extended *π*-conjugated molecules within a Self-Consistent Framework

Adam Coxson[1], Alessandro Troisi[1], Omer H. Omar[1]
[1]Chemistry Department, University of Liverpool, UK

acoxson@liverpool.ac.uk

π-conjugated molecules contain chains or rings of alternating single and double bonds which enable delocalisation of π electrons due to the overlap of adjacent p-type orbitals. Conjugated systems have many interesting optical and electronic properties that make them popular candidates for organic electronics, such as high charge transport Organic Semi-Conductors and π-conjugated polymers, which are vital for devices like Organic Photovoltaics (OPVs) and Organic Field-Effect Transistors (OFETs) [1]. π-conjugated materials are lightweight, low-cost, flexible, and have tuneable band-gaps, but such versatility comes at the cost of an extremely large design space which requires a combination of active-search and high-throughput virtual screening to study [2, 3].

The electronic structures of simple chain-like organic molecules consist of easily separable local environments, so they are not as rich nor diverse as their conjugate counterparts. For example, an α-amino acid called Isoleucine contains 3 of the most common terminating functional groups (RCOOH, CH3, NH2), as shown in Fig. 1. The electronic structures of such molecules are very localised and predictable; the bond lengths throughout the molecule will fall within a small range of values. In contrast, conjugated molecules are extremely delocalised due to the overlap of adjacent p-orbitals, which means that bond lengths and atomic interactions are not as easily quantifiable. For a machine learning algorithm to predict the electronic structure of Isoleucine, it only needs to learn the local structure of the terminal groups and how they are stitched together, as the terminal group and their connecting fragments will be very similar across different molecules. This simplifies the learning task and means that most 'unseen' test molecules containing these groups will lie within the interpolative domain space. On the other hand, delocalised conjugated systems demand a sufficiently large and diverse training set to cover their greater combinatorial space. However, even this is no guarantee for training success, as appropriate descriptors that can capture the intricacies of the delocalised electronic structure are required.

There are many organic molecule databases, with the QM7 and QM9 databases being the most commonly used benchmarks for machine learning applications [4]. However, the QM9 database only has small organic molecules of up to 9 heavy atoms of C, N, O, S and Cl, resulting in reduced chemical diversity and the lack of any extended π-conjugated systems. Furthermore, work by M. Glavatskikh et. al. has shown that QM9 has a particular lack of chemical diversity when compared to an equivalent dataset drawn from the PubChemQC database called PC9 [5].

The limitations of current datasets present a two-fold problem for machine learning: First, it restricts their generalizability, and second, it means ML struggles to predict larger and more complex molecules, which often have properties of interest. Consequently, there is a need for datasets that contain larger and more diverse molecules, such as those with extended π-conjugation. However, it is not sufficient to just apply current models to a more complex dataset, it also calls for the coincident development of physically informed machine learning descriptors that are capable of capturing as much system information as possible.

In this work, we have derived a dataset from the ZINC database that contains 13 million molecules that are commercially available [6]. We reduced this down to 150,000 molecules by clustering them according to unique conjugated cores, with a further reduction to 5000 molecules that contain between 10 to 25 C, N, O heavy atoms. This ensured that every molecule had a distinct electronic structure which is dictated by their conjugated core, despite the presence of any strongly interacting functional groups that may branch from the core.

Our aim is to use machine learning to predict the Fock matrix (one-electron Hamiltonian) of these conjugated cores to the typical accuracy of B3LYP Density Functional Theory (DFT), from an initial density guess. We do this by leveraging the self-consistent field (SCF) approach and have developed descriptors that use the overlap and density matrices in conjunction with an assembly of feed-forward networks to iteratively optimize the Fock matrix, as shown in Fig. 2. This replaces the computationally expensive step of estimating the Fock matrix from the previous density matrices using DFT. The matrix correlation heatmap in Fig. 3 shows how the overlap and density matrix block descriptors are effective at capturing the local information of the electronic structure, while embedding the model into an SCF framework allows it to account for non-locality.

By building a physically-informed machine learning model that integrates seamlessly with current SCF architecture, we can unite the interpolative power of machine learning with the interpretability and consistency of physical methods. One can also exploit the years of research put into improving convergence acceleration techniques for quantum chemical methods. Furthermore, curating datasets with ever greater diversity is an essential pre-requisite for developing models that can capture the complexity required to predict interesting candidates for organic electronic applications.

References

[1] Mateo-Alonso, A. “π-Conjugated materials: Here, there, and everywhere”. Chemistry of Materials 35.4 (2023): 1467-1469.

[2] Guo, Xin, Martin Baumgarten, and Klaus Müllen. "Designing π-conjugated polymers for organic electronics." *Progress in Polymer Science* 38.12 (2013): 1832-1908.

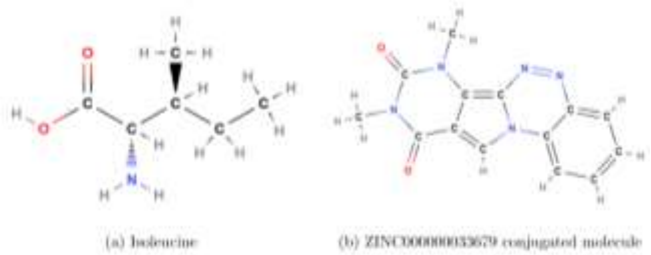
[3] Omar, Ö. H., Del Cueto, M., Nematiaram, T., & Troisi, A. (2021). “High-throughput virtual screening for organic electronics: a comparative study of alternative strategies”. *Journal of Materials Chemistry C*, *9*(39), 13557-13583.

[4] Ramakrishnan, R., Dral, P. O., Rupp, M., & Von Lilienfeld, O. A. (2014). “Quantum chemistry structures and properties of 134 kilo molecules”. Scientific data, 1(1), 1-7.

[5] Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T., & Da Mota, B. (2019). “Dataset’s chemical diversity limits the generalizability of machine learning predictions”. *Journal of Cheminformatics*, *11*, 1-15.

[6] Sterling, T., and John J. Irwin. “ZINC15–ligand discovery for everyone”. Journal of chemical information and modelling 55.11 (2015): 2324-2337.

Figures



(a) Isoleucine

(b) ZINC000000033679 conjugated molecule

Figure 1. (a) shows the Isoleucine molecule that consists of 2 methyl, 1 amino, and 1 carboxylic functional groups, all joined by a hydrocarbon chain. These components are very localised in electronic structure and ubiquitous across many similar molecules. (b) shows a conjugated molecule from the ZINC database (ZINC000000033679) with 4 adjacent rings of carbon and nitrogen over which π-electrons are delocalised, giving rise to more complex electronic structures. Due to their increased non-locality, attempts to develop descriptors for electronic structure are more difficult for (b)-type molecules.

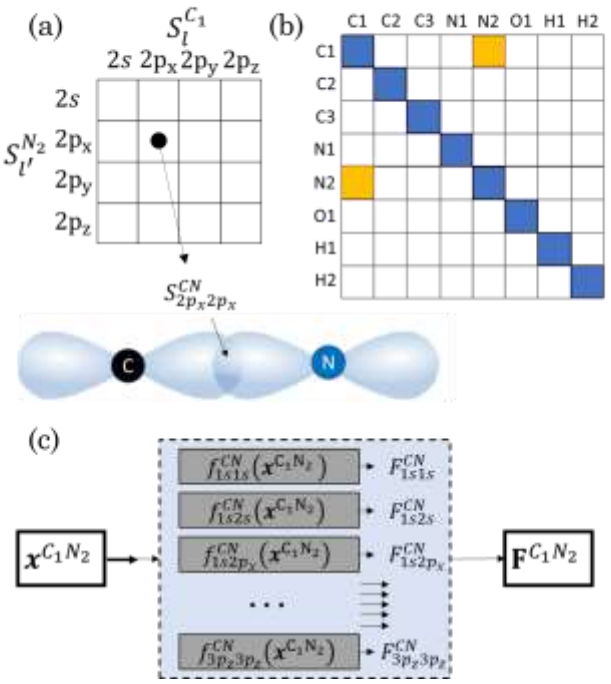


Figure 2. (a) The overlap of atomic orbitals between two atoms can be represented as a matrix block, which can then be built up into the full molecular overlap matrix (b). By breaking SCF matrices down into their atomwise blocks, an assembly of networks, as shown in (c), can be trained to predict individual matrix elements of the corresponding Fock blocks. This replaces the usually computationally expensive DFT step of estimating the Fock matrix from the density matrices.

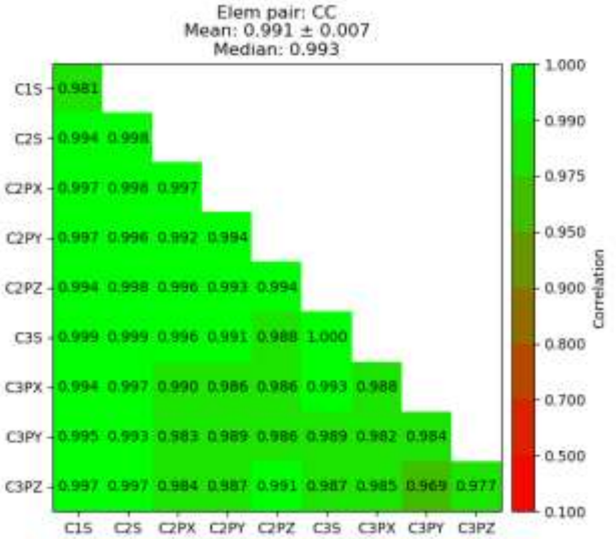


Figure 3. This shows the prediction results for the carbon-carbon pairwise interactions of a 500 molecule test set. For example, the C1s-C1s correlation shows how well one individual neural network performed in predicting C1s-C1s Fock matrix elements for the corresponding input density matrix block. This indicates how the density and overlap matrix block are effective descriptors of the local electronic structure.