Communications in Computational Chemistry
Commun. Comput. Chem. Vol. 8, No. 1, 37–43
doi: 10.4208/cicc.2025.358.01

Regular Article

VBData: A Valence Bond Property Dataset of Single Bonds for Organic Compounds

Mingwei Yang[1], Yameng Zheng[1], Muhammad Shoaib[1], Wei Wu[2,*] and Jinshuai Song[1,*]

[1] *College of Chemistry, Pingyuan Laboratory, and State Key Laboratory of Cotton Bio-breeding and Integrated Utilization, Zhengzhou University, Zhengzhou, Henan 450001, China;*

[2] *State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, Fujian 361005, China.*

[*] Corresponding authors: weiwu@xmu.edu.cn; jssong@zzu.edu.cn.

Received 24 Dec. 2025; Accepted (in revised version) 17 Feb. 2026

Abstract: The properties of valence bonds are fundamental to understand and predict the reactivity of chemical reactions. This article presents a comprehensive dataset derived from quantum chemical calculations performed at a consistent computational level for more than 40,000 organic compounds (containing only C, H, N, and O atoms) and over 400,000 non-cyclic single bonds within them. Key bond properties include resonance energies and structural weights. The valence bond computation was performed using XMVB program at the VBSCF level, based on the equilibrium geometries optimized at the B3LYP/def2-SVP level. This dataset is anticipated to serve as a new standard benchmark for bond properties in organic chemistry and to provide a foundation for developing machine learning models.

Key words: bond dataset, valence bond theory, resonance energy, VBData, structural weights.

1. Introduction

Chemical bond properties are significant in predicting the chemical reactivity and selectivity of chemical reactions [1-4]. Key descriptors such as bond length, bond order, bond dissociation energy (BDE), and acid dissociation constant (pKa) could be obtained from thermodynamic experiments or quantum chemical calculations. Among these, valence bond (VB) properties serve as crucial indicators of molecular kinetic stability [5-7], aromaticity [8-10], magnetic behavior [11], and chemical reactivity [12,13]. Compared to thermochemical experiments and quantum chemistry based on molecular orbital based quantum chemistry, valence bond (VB) theory offers superior intuitive insights into chemical bonding. It provides a clear physical explanation for bond strength and directionality, which facilitates the understanding of molecular stability and chemical reactions. Traditional quantum chemical methods based on valence bond theory can accurately predict diverse molecular properties but require high computational cost. These limitations consequently restrict the application of VB methods to rapid screening of functional molecules in realistic chemical systems.

The prediction of bond properties could be achieved by machine learning models, driven by the continuous advancement of artificial intelligence [14-17]. Machine learning can uncover latent relationships between feature descriptors and properties, overcoming the limitations of conventional methods. Machine learning is extensively applied across domains, benefiting from its rapid processing and user-friendly nature. These machine learning models rely on corresponding chemical property databases. However, existing datasets still exhibit critical gap in coverage and applicability. To date, several high-quality chemical datasets (e.g., ioChem-BD [18], BDE-db [19,20], iBOND, BDNCM [1], and the QM series [21,22]) have been developed and made publicly available to support machine learning models and research in chemical property prediction. Nevertheless, most datasets are highly specialized and thus unsuitable for general use. Additionally, the lack of high-quality datasets to support model training has left the domain of valence bond property largely unexplored.

In modern valence bond theory, the description of a single bond between atoms A and B requires a superposition of multiple resonance structures, as shown in **Figure 1**.

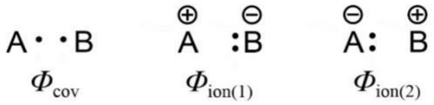

Figure 1. Three resonance structures of a single bond.

As shown in Eq. (1), the VB wavefunction Ψ is expressed as a combination of the covalent form Φ_{cov}(A·-·B) and two ionic forms, Φ_{ion}(A^+B^-) and Φ'_{ion} (A^-B^+) [23,24]:

$$\Psi = C_1\Phi_{\mathrm{cov}} + C_2\Phi_{\mathrm{ion}} + C_3\Phi_{\mathrm{ion}}^{'} \qquad (1)$$

The resonance energy (RE) [25,26] can be determined via thermochemical methods [27], hydrogenation heat experiments, or quantum chemical calculations [28]. However, experimental determination of resonance energy for polyatomic molecules remains technically challenging and cost-prohibitive. The resonance energy [29] is defined as the energy difference between the lowest-energy structure and the full wavefunction described by Eq. (1). For the small organic molecules studied in this work, the covalent structure exhibits a relatively low energy. Thus, the single-bond RE is calculated as Eq. (2) [30]:

$$RE = E(\Phi_{\mathrm{cov}}) - E_{\Psi} \qquad (2)$$

where E_{Ψ} is the energy of the total wavefunction (Ψ) and $E(\Phi_{\mathrm{cov}})$ is the energy of the most stable structure, which represents the energy of the covalent structure in this study.

The overall stability arises from a superposition of resonance structures, each assigned a specific weight that quantifies its contribution (**Figure 1**). The normalized structural weight in Valence Bond Self-Consistent Field (VBSCF) [31] method is defined as [32]:

$$W_K = C_K\langle\Phi_K|\Psi\rangle = \sum_{L} C_K M_{KL} C_L \qquad (3)$$

where W_K denotes the structural weight, C_K and C_L are the structural coefficients in the wavefunction, and M_{KL} represents overlap matrices between K and L structures. This study focuses on nine key chemical bond types (C-H, C-C, H-N, C-N, C-O, H-O, N-O, N-N and O-O), all of which are suitable for computational simulation with the present VB model [30].

In this article, we report a dataset called VBData including the valence bond properties of single bonds for organic compounds. The dataset has 40,837 organic compounds and 400,972 chemical bonds, with systematic documentation of their valence bond properties. All molecular geometries were optimized at the B3LYP/def2-SVP theoretical level [33,34], following bond property calculations by using the VBSCF [31] method. This computational protocol was carefully selected to achieve an optimal balance between accuracy and computational efficiency. The multidimensional information effectively captures both the bond strength characteristics and chemical reactivity profiles of the investigated bonds. We expect graph neural network (GNN) [35] models to be applied for valence bond property prediction based on this dataset.

2. Theoretical method

2.1 Molecular and chemical bond selection criteria.

The molecular structures were sourced from the same PubChem Compound database [36] utilized by BDE-db. Selection criteria included neutral organic molecules containing fewer than 10 heavy atoms (excluding hydrogen) and composed exclusively of C, H, O, N. All SMILES strings were standardized and deduplicated using RDKit [37].

2.2 Conformational optimization.

Initial molecular geometries were generated from SMILES strings using RDKit. Notably, while the default 3D conformation generation excludes hydrogen atoms, we explicitly added hydrogen atoms prior to optimization due to their critical influence on molecular conformation. To enhance the quality of 3D conformers, we adopted a hierarchical workflow as shown in **Figure 2**. Firstly, 500 conformers for every molecule were generated using the ETKDG method [38]. Next, these conformers were optimized by the MMFF94 force field [39]. Subsequently, the lowest-energy conformation was selected as the initial geometry for subsequent high-level optimizations. The obtained geometry was exported to an XYZ file and further optimized by using the semi-empirical ODM3 [40] method within the MNDO [41] quantum chemical program to improve structural accuracy and reduce computational overhead in later stages. Final geometry was optimized at the B3LYP/def2-SVP level in Gaussian 16 [42] (computational details of structure optimization are provided in the Supporting Information S1-S2). All Gaussian optimized structures were rigorously verified for energy convergence.

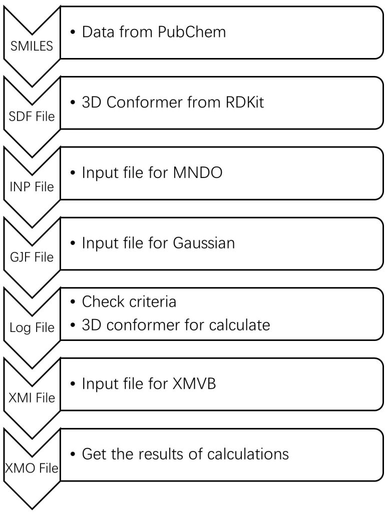

Figure 2. Overview of the workflow for this work, including the software and files utilized.

2.3 Valence bond calculations.

To maximize the diversity of bond types per unit of computational effort and avoid complex intramolecular interactions in strained rings, such as ring strain, non-cyclic single bonds that are not in rings were selected from the optimized molecular structures for data collection, to enable the systematic construction of a comprehensive dataset. The valence-bond properties of bonds in rings are strongly influenced by the ring structure and the local chemical environment, resulting in a broad distribution that poses significant challenges for effective sampling. These bonds were calculated by using the XMVB program [43,44], to collect structural weights and resonance energy. The computational procedures are depicted in **Figure 2**. For each targeted bond, two separate XMVB calculations were performed to compute E_{Ψ} by 3 structures and $E(\Phi_{\mathrm{cov}})$ by the covalent structure. The active space was defined with 2 active orbitals in 2 active