EBOND: AN ELECTRON-GUIDED FRAGMENTATION DATASET FOR MOLECULAR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular fragmentation, which decomposes molecules into smaller substructures, is widely used in fragment-based molecular learning. Existing fragmentation schemes primarily determine cleavage sites using hand-crafted chemical rules or cleavage patterns derived from reaction data. These approaches produce fragment spaces shaped by pre-defined structural assumptions or the coverage of recorded reactions, while largely overlooking molecule-specific electronic structure. Electronic structure underlies molecular properties, motivating its use in molecular fragmentation. However, obtaining such electronic information at scale requires costly quantum-chemical calculations, limiting the development of large-scale electron-guided fragmentation datasets. In this work, we introduce **EBOND**, a large-scale quantum-chemical dataset for electron-guided molecular fragmentation, covering 452,775 molecules at a computational cost of approximately 1.26 million CPU core-hours. Using three-state density functional theory and wave-function analysis, EBOND provides atom- and bond-level electronic descriptors and translates them into molecule-specific scores for fragmentation. On USPTO-50K, EBOND-derived cuts cover 81.1% of recorded synthetic disconnections and recover 73.4% of those missed by both BRICS and RECAP. We further show that EBOND-derived fragments provide useful learning signals for retrosynthesis and fragment-based molecular optimization. These results demonstrate the value of molecule-specific electronic information for molecular fragmentation and establish EBOND as a reusable resource for fragment-based molecular learning. Code are available at `https://anonymous.4open.science/r/EBOND`.

1 INTRODUCTION

Molecular fragments are widely used as intermediate representations in data-driven molecular learning (Erlanson et al., 2004), supporting tasks such as molecular generation (Jin et al., 2018; Lee et al., 2024), molecular optimization (Xie et al., 2021; Kaech et al., 2026), and synthesis planning (Shee et al., 2026; Chen et al., 2026). These substructures serve as the basic units of fragment-based models, so the choice of fragmentation boundaries directly determines how a molecule is represented (Jinsong et al., 2024; Xu & Kang, 2025).

Existing methods define fragmentation boundaries mainly through pre-defined chemical rules, reaction precedents, or data-driven patterns. Rule-based methods, such as RECAP (Lewell et al., 1998) and BRICS (Degen et al., 2008), apply fixed cutting rules derived from chemical knowledge, limiting fragmentation to bond environments covered by these rules. Reaction-derived methods infer disconnections from transformations recorded in large reaction datasets (Coley et al., 2017), but depend on the coverage and bias of available reaction data. Recent data-driven methods learn molecular representations to identify fragments suited to particular tasks (Yang et al., 2024; Zhang et al., 2025). Despite this progression, fragmentation decisions are still made mainly from structural patterns or observed reactions, while overlooking molecular electronic structure that may provide a molecule-specific signal for defining fragmentation boundaries (Figure 1).

Chemical bonds arise from electronic interactions between atoms (Bader, 1985), making electronic information useful for describing differences among bonds in a molecule. Since fragmentation boundaries are defined through bond selection, such information provides molecule-specific sig-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nals beyond local structural patterns. Electronic responses to electron addition and removal further capture changes in local electronic structure at the atom and bond levels (Geerlings et al., 2003; Gonzalez-Suarez et al., 2012). However, obtaining such information at scale requires costly quantum-chemical calculations, making studies of electron-guided fragmentation challenging.

To this end, we introduce **EBOND**, a large-scale quantum-chemical dataset for electron-guided molecular fragmentation. EBOND comprises 452,775 molecules with three-state DFT calculations followed by wave-function analysis, providing atom- and bond-level electronic descriptors. Using these descriptors, we further construct molecule-specific electron-guided fragmentation annotations that translate electronic information into fragmentation boundaries. On USPTO-50K, EBOND-derived cuts cover 81.1% of recorded synthetic disconnections and recover 73.4% of those missed by both BRICS and RECAP, demonstrating their synthetic relevance and complementarity to rule-based methods. EBOND-derived fragments further support retrosynthesis and fragment-based molecular optimization, demonstrating the broader utility of EBOND as a resource for fragment-based molecular learning. Our contributions are summarized as follows:

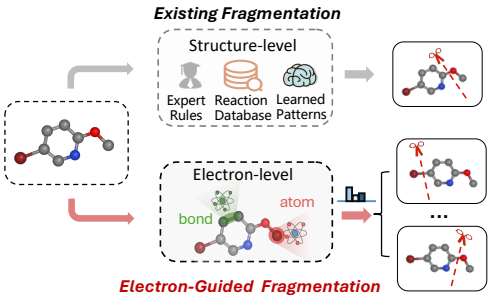


Figure 1: Overview of existing fragmentation schemes and electron-guided EBOND.

- We introduce **EBOND**, a large-scale quantum-chemical dataset that provides atom- and bond-level electronic information for 452,775 molecules;
- Using these electronic descriptors, we further construct molecule-specific electron-guided fragmentation annotations that identify fragmentation boundaries;
- Extensive analyses validate the effectiveness and utility of EBOND-derived fragmentation, demonstrating its synthetic relevance and complementarity to rule-based methods, as well as benefits for retrosynthesis and molecular optimization.

2 Background and Related Work

2.1 Molecular Fragmentation Algorithms

Molecular fragmentation decomposes molecules into smaller substructures by cutting selected bonds, and the resulting fragments serve as basic units for molecular generation, optimization, and synthesis planning. Existing methods determine fragmentation boundaries using chemical rules, reaction data, or learned patterns. Rule-based methods, including RECAP (Lewell et al., 1998) and BRICS (Degen et al., 2008), define cleavage sites through pre-defined bond environments, restricting fragmentation to bonds covered by these rules. Later variants broaden fragment coverage through fragment enumeration and additional cutting rules (Zabolotna et al., 2021; Diao et al., 2023; Zhang et al., 2024). Reaction-derived approaches identify fragmentation boundaries from bond changes or templates extracted from recorded reactions (Coley et al., 2017), but the resulting patterns depend on the coverage and biases of available reaction data. More recent data-driven methods, such as DigFrag (Yang et al., 2024), learn task-specific fragmentation patterns from molecular graphs, but their learned boundaries depend on the training data and task objective. In contrast, EBOND introduces an electronic-structure perspective to molecular fragmentation by defining boundaries from molecule-specific bond-level information.

2.2 Electronic Structure and Quantum Chemistry Datasets

Density functional theory (DFT) provides a practical description of molecular electronic structure through the electron density. According to the Hohenberg–Kohn theorem, the ground-state density determines the external potential and hence the ground-state properties of the system, while Kohn–Sham DFT provides a practical framework for obtaining this density (Hohenberg & Kohn, 1964; Kohn & Sham, 1965). Beyond ground-state properties, conceptual DFT describes how the electron

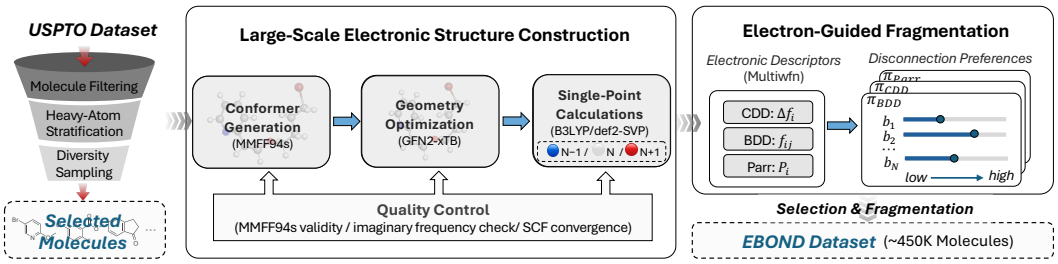


Figure 2: **Construction of EBOND Dataset.** USPTO molecules are filtered and sampled, followed by electronic-structure calculations and wave-function analysis. The resulting electronic descriptors guide molecular fragmentation to construct EBOND.

density changes with electron number at fixed external potential (Parr & Yang, 1984; Geerlings et al., 2003). These changes form the basis of local reactivity descriptors and can be described at the atomic level through population analysis and at the bond level through changes in bond order (Yang & Mortier, 1986; Gonzalez-Suarez et al., 2012; Pino-Rios, 2025). Further theoretical background is provided in Appendix A.

Quantum-chemical datasets have developed rapidly in both scale and the electronic information they provide. QM9 provides equilibrium molecular structures together with a range of ground-state quantum-chemical properties (Ramakrishnan et al., 2014). QMugs extends quantum-chemical calculations to a large collection of drug-like molecules (Isert et al., 2022), QH9 provides molecular Hamiltonian matrices (Yu et al., 2023), and EDBench provides large-scale electron-density data (Xiang et al., 2026). These resources mainly describe molecular electronic structure itself, while changes with electron number have received less attention. In contrast, EBOND connects quantum-chemical information with molecular fragmentation by deriving atom- and bond-level electronic descriptors and translating them into molecule-specific fragmentation scores and annotations.

3 The EBOND Dataset

3.1 Construction of the EBOND Dataset

We construct EBOND by performing large-scale three-state DFT calculations on 538,998 molecules sampled from the USPTO dataset (Lowe, 2012), following the pipeline shown in Figure 2. For each molecule, we generate conformers with RDKit (Landrum et al., 2022), minimize and rank them using the MMFF94s force field (Halgren, 1996; 1999), and select the lowest-energy conformer. The selected conformer is then optimized at the GFN2-xTB level (Bannwarth et al., 2019). At the optimized geometry, we perform single-point calculations for the neutral N-electron state and the corresponding $N \pm 1$-electron states at the B3LYP/def2-SVP level (Becke, 1993; Lee et al., 1988; Weigend & Ahlrichs, 2005) using Gaussian 16 (Frisch et al., 2016). All three states are evaluated at the same geometry without further structural relaxation. The neutral state is treated as a singlet using restricted Kohn–Sham DFT, while the charged states are treated as doublets using unrestricted Kohn–Sham DFT. For molecules for which all three calculations converge, the wavefunctions are analyzed with Multiwfn (Lu & Chen, 2012; 2022; Lu, 2024) to obtain atom- and bond-level electronic descriptors. Only molecules that pass the quality-control checks at all stages are included, resulting in 452,775 molecules in the final dataset. All calculations were performed on a high-performance computing cluster comprising 315 compute nodes, each equipped with two AMD EPYC 7452 32-core processors and 64 physical CPU cores, for a total of 20,160 physical CPU cores. Constructing EBOND required approximately 1.26 million CPU core-hours, with three-state DFT calculations accounting for 91.5% of the total cost. Further construction details are provided in Appendix B.

3.2 Electronic Descriptors

We represent each molecule as an undirected graph $\mathcal{M} = (\mathbb{V}, \mathbb{B})$, where $\mathbb{V}$ and $\mathbb{B}$ denote the sets of atoms and covalent bonds, respectively. Each bond $b_{ij} \in \mathbb{B}$ connects atoms i and j. Using the N, $N+1$, and $N-1$ electronic states calculated at the same molecular geometry, EBOND de-

rives three descriptors of local electronic response. The condensed dual descriptor (CDD) (Morell et al., 2005) describes the atomic response to electron addition and removal, the bond dual descriptor (BDD) (Pino-Rios, 2025) describes the corresponding response at the bond level, and Parr functions (Domingo et al., 2013) describe local electrophilic and nucleophilic character from spin density. These descriptors are mapped to bond-level disconnection scores in Section 3.3.

Condensed dual descriptor. At the atomic level, the responses to electron addition and removal are described by the condensed Fukui functions (Parr & Yang, 1984; Yang & Mortier, 1986; Oláh et al., 2002). Using the Hirshfeld atomic charge $q_i^{(n)}$ of atom i in the n-electron state, the condensed Fukui functions are defined as $f_i^+ = q_i^{(N)} - q_i^{(N+1)}$, and $f_i^- = q_i^{(N-1)} - q_i^{(N)}$. Their difference gives the condensed dual descriptor (Morell et al., 2005), $\mathrm{CDD}_i = \Delta f_i = f_i^+ - f_i^- = 2q_i^{(N)} - q_i^{(N+1)} - q_i^{(N-1)}$. Positive and negative values indicate electrophilic and nucleophilic character, respectively.

Bond dual descriptor. Bond-level electronic response can be described through changes in bond order upon electron addition and removal (Gonzalez-Suarez et al., 2012). For bond b_{ij}, the bond order in the n-electron state is denoted by $\mathrm{BO}_{ij}^{(n)}$ (Mayer & Salvador, 2004). The corresponding bond Fukui functions are $f_{ij}^+ = \mathrm{BO}_{ij}^{(N+1)} - \mathrm{BO}_{ij}^{(N)}$ and $f_{ij}^- = \mathrm{BO}_{ij}^{(N)} - \mathrm{BO}_{ij}^{(N-1)}$. Based on these responses, the bond dual descriptor (BDD) is defined as $\mathrm{BDD}_{ij} = f_{ij}^+ - f_{ij}^- = \mathrm{BO}_{ij}^{(N+1)} - 2\mathrm{BO}_{ij}^{(N)} + \mathrm{BO}_{ij}^{(N-1)}$ (Pino-Rios, 2025). Positive and negative BDD values indicate electrophilic and nucleophilic bond responses, respectively.

Parr functions. Parr functions describe local electrophilic and nucleophilic character from the spin density of the charged states (Domingo et al., 2013). Using the Hirshfeld atomic spin populations, the electrophilic and nucleophilic Parr functions for atom i are defined as $P_i^+ = s_i^{(N+1)}$ and $P_i^- = s_i^{(N-1)}$, respectively, where $s_i^{(n)}$ denotes the spin population of atom i in the n-electron state.

3.3 Electron-guided Fragmentation

Electronic descriptors provide continuous signals on atoms and bonds, whereas molecular fragmentation requires discrete decisions over bonds. eBOND therefore converts these signals into bond-level fragmentation scores. In this work, we restrict fragmentation to non-ring heavy-atom bonds, denoted by $\mathcal{B}(\mathcal{M})$. For BDD, the positive and negative responses are separated into electrophilic and nucleophilic channels as $\pi_{\mathrm{BDD}}^{\mathrm{elec}}(b_{ij}) = \max(\mathrm{BDD}_{ij}, 0)$ and $\pi_{\mathrm{BDD}}^{\mathrm{nuc}}(b_{ij}) = \max(-\mathrm{BDD}_{ij}, 0)$. For CDD, we consider bonds whose two atoms have opposite signs and define $\pi_{\mathrm{CDD}}(b_{ij}) = |\mathrm{CDD}_i - \mathrm{CDD}_j|$. For Parr functions, the local electrophilicity and nucleophilicity are defined as $\omega_i = \omega \max(P_i^+, 0)$ and $\mathcal{N}_i = \mathcal{N} \max(P_i^-, 0)$, where ω and $\mathcal{N}$ are the molecular electrophilicity and nucleophilicity obtained from the conceptual DFT analysis in Multiwfn, respectively. The corresponding bond score is $\pi_{\mathrm{Parr}}(b_{ij}) = \max(\omega_i \mathcal{N}_j, \omega_j \mathcal{N}_i)$, using the stronger electrophile–nucleophile pairing between the two atoms. All scores are defined so that larger values indicate stronger electronic support for fragmentation. Since the scores are derived from molecule-specific electronic responses, bonds with the same local structure can be assigned different scores.

To translate these scores into discrete fragmentation decisions, eBOND evaluates each bond relative to the other candidate bonds in the same molecule. For each fragmentation score π, the scores are standardized within the molecule as $z_\pi(b_{ij}) = [\pi(b_{ij}) - \mu_\pi(\mathcal{M})]/\sigma_\pi(\mathcal{M})$, where $\mu_\pi(\mathcal{M})$ and $\sigma_\pi(\mathcal{M})$ are the corresponding mean and standard deviation. This within-molecule normalization measures the electronic support for each bond relative to other bonds in the same molecule and allows a shared threshold $\tau = 0.25$ to be applied across molecules and descriptors. A bond $b_{ij} \in \mathcal{B}(\mathcal{M})$ is selected when $z_\pi(b_{ij}) \geq \tau$ for at least one fragmentation score. The union of the selected bonds defines $\mathcal{B}_{\mathrm{eBOND}}(\mathcal{M})$. These bonds are then removed simultaneously, and hydrogen atoms are added to the resulting open valences to generate the fragments. Further details on threshold selection are provided in Appendix E.2.

3.4 Dataset Statistics

The final eBOND dataset contains 452,775 molecules and provides bond-level electronic information for approximately 12.1 million heavy-atom bonds. Electron-guided fragmentation selects

approximately 1.87 million fragmentation boundaries, producing approximately 154,000 unique fragments. Figure 3 summarizes the molecular, bond, and fragmentation statistics of the dataset. Approximately 80% of the molecules contain 13–37 heavy atoms, with a median of 24 (Figure 3(a)). The dataset contains 10 non-hydrogen elements and 67 heavy-atom bond types, with the ten most frequent bond types accounting for 95.0% of all heavy-atom bonds (Figure 3(b)). Figure 3(c) further shows the distribution of the number of selected bonds per molecule. Additional statistics are provided in Appendix C.

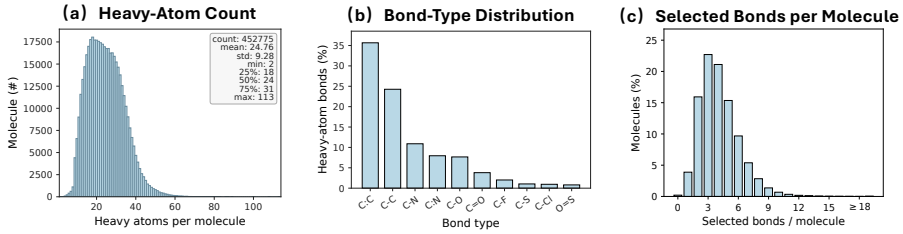


Figure 3: **Statistics of the EBOND dataset.** (a) Distribution of heavy-atom counts across molecules. (b) Distribution of the ten most frequent heavy-atom bond types. (c) Distribution of the number of selected bonds per molecule.

4 Data Analysis

Molecular fragmentation defines the substructures used to represent a molecule, making the choice of fragmentation boundaries critical. We evaluate electron-guided fragmentation through three questions: *(i)* whether electronic response provides information beyond local structural patterns (Section 4.2), *(ii)* whether the resulting fragmentation aligns with recorded synthetic disconnections (Section 4.3), and *(iii)* how the resulting fragment space differs from rule-based fragmentation and whether these fragments are represented among commercial building blocks (Section 4.4).

4.1 Evaluation Settings

Evaluation data. We conduct our analyses on USPTO-50K, denoted as $\mathcal{D}_{\mathrm{rxn}}$, which is derived from the USPTO reaction dataset (Lowe, 2012), the same data source used to construct EBOND. Product-side bond-disconnection labels are extracted following Somnath et al. (2021) and are used only for the synthetic relevance analysis. We additionally use the standardized Enamine US Stock collection of 282,727 compounds to assess building-block coverage (Shivanyuk et al., 2007).

Baselines. We compare EBOND with random fragmentation, BRICS (Degen et al., 2008), and RECAP (Lewell et al., 1998). For each molecule, the random baseline randomly selects the same number of non-ring heavy-atom bonds as EBOND, and results are averaged over 10 random seeds. BRICS and RECAP are applied using their original fragmentation rules.

Evaluation metrics. Molecular fragmentation does not have a unique ground truth, so we evaluate EBOND from complementary perspectives. We use **recorded-disconnection coverage (RDC)** to measure agreement with recorded synthetic disconnections, **missed-disconnection recovery (MDR)** to measure complementarity with rule-based fragmentation, and **building-block coverage (BCR)** to measure the representation of the resulting fragments in commercial building blocks. For a molecule M, $\mathcal{B}(M)$ denotes the set of non-ring heavy-atom bonds considered for fragmentation, and $\mathcal{B}_m(M)$ denotes the bonds selected by method m. For each $(M, b) \in \mathcal{D}_{\mathrm{rxn}}$, b denotes the recorded disconnection, and RDC is defined as $\mathrm{RDC}_m = |\{(M, b) \in \mathcal{D}_{\mathrm{rxn}} \mid b \in \mathcal{B}_m(M)\}|/|\mathcal{D}_{\mathrm{rxn}}|$. For a rule-based reference method r, $\mathcal{D}^r_{\mathrm{miss}} = \{(M, b) \in \mathcal{D}_{\mathrm{rxn}} \mid b \in \mathcal{B}(M),\, b \notin \mathcal{B}_r(M)\}$ denotes the recorded disconnections within the fragmentation scope that are missed by r, and MDR is defined as $\mathrm{MDR}_r = |\{(M, b) \in \mathcal{D}^r_{\mathrm{miss}} \mid b \in \mathcal{B}_{\mathrm{eBOND}}(M)\}|/|\mathcal{D}^r_{\mathrm{miss}}|$. For method m, $\mathcal{V}_m$ denotes its attachment-aware fragment vocabulary, and BCR is defined as $\mathrm{BCR}_m = |\{u \in \mathcal{V}_m \mid \exists e \in \mathcal{E},\, u \preceq e\}|/|\mathcal{V}_m|$, where $\mathcal{E}$ denotes the Enamine building-block collection and $u \preceq e$ indicates that fragment u matches a substructure of e. Further evaluation details are provided in Appendix D.

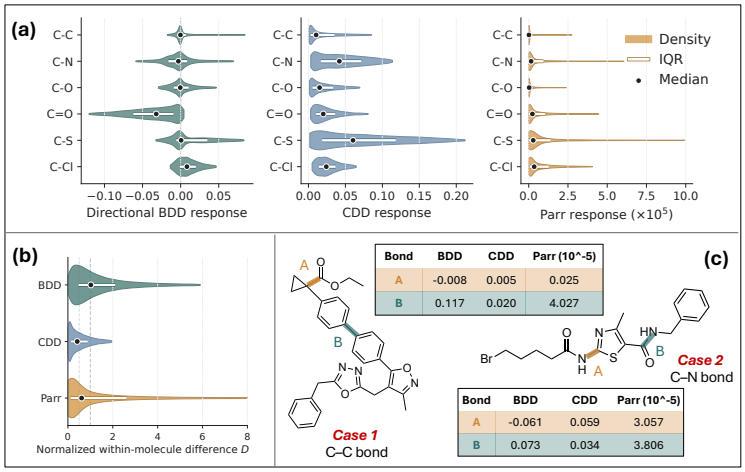


Figure 4: **Variation in electronic responses.** (a) Response distributions across common bond types. (b) Variation among bonds with the same radius-2 local environment within a molecule. (c) Representative C–C and C–N pairs with different electronic responses.

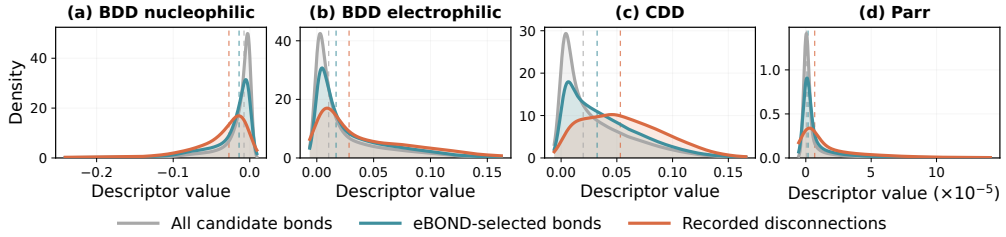


Figure 5: **Electronic responses of selected bonds and recorded disconnections.** Distributions of all candidate bonds, EBOND-selected bonds, and recorded disconnections for (a) nucleophilic BDD, (b) electrophilic BDD, (c) CDD, and (d) Parr. Dashed lines indicate the medians.

4.2 Electronic Information Beyond Local Structure

Existing fragmentation methods mainly define cleavage boundaries from local structural patterns. We examine whether electronic information provides additional bond-level signals beyond these patterns. We first compare directional BDD responses, which distinguish nucleophilic and electrophilic directions, together with CDD and Parr responses across common bond types. Figure 4a shows variation in electronic responses among bonds of the same type. We then perform a stricter comparison by considering bonds with the same radius-2 local environment within the same molecule. For each group g of such bonds and electronic descriptor d, we measure the variation as $D_{g,d} = \text{median}_{i<j} |x_{i,d} - x_{j,d}| / \text{IQR}(x_d^{\text{bond type}})$, where $x_{i,d}$ is the response value of bond i for descriptor d, and IQR is the interquartile range of the corresponding descriptor over all bonds of the same type. Figure 4b shows that the variation remains even when the local structural environment is the same, with median D values of 1.02, 0.42, and 0.61 for BDD, CDD, and Parr, respectively. Representative C–C and C–N bond pairs further illustrate differences in electronic responses among bonds of the same type within the same molecule (Figure 4c). These results show that bonds with the same local structure can exhibit different electronic responses, supporting the use of molecule-specific electronic information for defining fragmentation boundaries.

4.3 Synthetic Relevance of Electron-Guided Fragmentation

We then examine the synthetic relevance of EBOND by asking whether selected bonds show similar electronic patterns to known disconnections, how often they recover these disconnections, and whether they recover disconnections missed by rule-based methods.

Electronic-response alignment. We compare the electronic response distributions of all candidate bonds, EBOND-selected bonds, and recorded disconnections. As shown in Figure 5, both EBOND-selected bonds and recorded disconnections shift in the same direction relative to all candidate bonds across nucleophilic BDD, electrophilic BDD, CDD, and Parr responses. For example, the median CDD response increases from 1.98×10^{-2} for all candidate bonds to 3.23×10^{-2} for EBOND-selected bonds and 5.32×10^{-2} for recorded disconnections. Similar trends are observed for the other electronic responses. These results support the use of molecule-specific electronic responses as synthesis-relevant signals for defining fragmentation boundaries.

Table 1: **Disconnection coverage.** RDC is the recovered fraction of recorded disconnections. Random uses the same cut number as EBOND.

Method	Avg. Cuts / mol	RDC (%)
Random	6.17	54.4
RECAP	1.83	58.1
BRICS	5.07	82.3
EBOND	6.17	81.1

Table 2: **EBOND complementarity.** MDR is the missed-disconnection recovery rate. C-RDC is the combined coverage with EBOND.

Base method	MDR (%)		C-RDC (%)
	Random	EBOND	+ EBOND
RECAP	53.0	**75.0**	89.5
BRICS	51.2	**75.8**	95.7
BRICS ∪ RECAP	51.0	**73.4**	**96.1**

Recorded-disconnection coverage. We then evaluate synthetic relevance by measuring the recovery rate of recorded synthetic disconnections. On the reaction evaluation set $\mathcal{D}_{\mathrm{rxn}}$, EBOND achieves an RDC of 81.1% with an average of 6.17 cuts per molecule, without using reaction labels or predefined cleavage rules (Table 1). BRICS and RECAP achieve RDC values of 82.3% and 58.1%, respectively. With the same number of cuts, the random baseline achieves an RDC of 54.4%. These results show that EBOND-guided fragmentation captures synthesis-relevant disconnection patterns.

Complementarity with rule-based fragmentation. We further examine whether EBOND recovers recorded disconnections missed by BRICS and RECAP. As shown in Table 2, EBOND recovers 75.0% of the disconnections missed by RECAP and 75.8% of those missed by BRICS. When combined with RECAP and BRICS, EBOND achieves C-RDC values of 89.5% and 95.7%, respectively. For disconnections missed by both BRICS and RECAP, EBOND recovers 73.4%, compared with 51.0% from randomly selecting the same number of additional bonds, and increases C-RDC from 85.3% to 96.1%. These results show that molecule-specific electronic information captures synthesis-relevant disconnection patterns beyond those encoded by rule-based fragmentation.

4.4 Fragment-Space Analysis

We further analyze the fragment space in terms of fragment composition, physicochemical properties, and commercial building-block coverage.

Fragment composition. Following the fragmentation method described in Section 3.3, EBOND, BRICS, and RECAP produce 26,168, 14,930, and 34,600 unique fragments, respectively. We first compare the sizes of the resulting fragments. As shown in Figure 6(a), the three methods cover similar fragment-size ranges, with median HAC values of 12 for EBOND, 11 for BRICS, and 13 for RECAP. We then compare the fragments within individual HAC groups. Figure 6(b) shows that EBOND generates both shared and method-specific fragments within the same HAC range. For $3 \leq \mathrm{HAC} \leq 10$, 57.0% of EBOND fragments are unique to EBOND. These results show that EBOND generates distinct fragment structures compared with rule-based methods while maintaining similar overall fragment sizes.

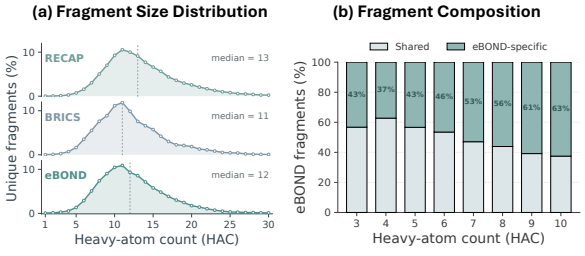


Figure 6: (a) Fragment-size distributions of EBOND and rule-based methods. (b) Shared and EBOND-specific fragments across HAC groups.

Fragment properties. Beyond fragment-space composition, we compare the properties of fragments generated by the three methods. As shown in Table 3, EBOND fragments show higher Fsp3 (0.51 vs. 0.34 and 0.37) and lower aromaticity (0.39 vs. 0.48 and 0.47) than BRICS and RECAP. These

Table 3: **Fragment properties.** Values are reported as mean ± standard deviation over fragments.

Method	MW	cLogP	TPSA	RotB	Fsp3	Aromatic frac.	Rings	Heteroatom frac.
BRICS	179.42 ± 65.09	1.67 ± 1.40	37.78 ± 22.95	0.93 ± 1.71	0.34 ± 0.33	0.48 ± 0.32	1.76 ± 1.07	0.28 ± 0.14
RECAP	200.30 ± 73.18	1.98 ± 1.47	40.59 ± 22.68	2.01 ± 1.95	0.37 ± 0.29	0.47 ± 0.26	1.72 ± 0.97	0.26 ± 0.12
EBOND	176.27 ± 60.22	1.61 ± 1.40	31.71 ± 19.13	1.89 ± 1.84	0.51 ± 0.33	0.39 ± 0.31	1.69 ± 1.04	0.25 ± 0.13

trends remain when the comparison is restricted to fragments with $3 \leq \text{HAC} \leq 10$. These results show that EBOND generates fragments with different physicochemical properties from rule-based methods, providing complementary structural units for fragment-based learning.

Fragment frequency and building-block coverage. We further examine the commercial building-block coverage (BCR) of the resulting fragment space. BCR is computed using attachment-aware fragments that preserve their original connection sites. For a fragment u, $r(u)$ denotes its frequency in the dataset. As shown in Table 4, EBOND achieves a BCR of 50.7% over all fragments. For method-specific fragments, EBOND achieves a BCR of 43.1%, compared with 41.4% for BRICS and 31.9% for RECAP. The BCR further increases to 70.1% and 84.3% for EBOND-specific fragments occurring in at least two and five molecules, respectively. These results show that EBOND expands the fragment space while maintaining strong support from commercially available building blocks, providing additional fragments that remain practical for fragment-based molecular design.

Table 4: **Building-block coverage.** BCR (%, higher is better).

Fragment subset	BRICS	RECAP	**eBOND**
All fragments			
All	**51.2**	41.6	50.7
$r(u) \geq 2$	77.3	72.0	**79.2**
$r(u) \geq 5$	90.8	86.6	**92.2**
Method-specific fragments			
All	41.4	31.9	**43.1**
$r(u) \geq 2$	62.7	58.0	**70.1**
$r(u) \geq 5$	74.2	70.3	**84.3**

5 Downstream Applications

5.1 Retrosynthesis

We evaluate EBOND on single-step retrosynthesis using a pre-training–fine-tuning setting. Pre-training pseudo-reactions are constructed by applying random, BRICS, and EBOND fragmentation exclusively to product molecules from the USPTO-50K training split, with validation and test products excluded. Random fragmentation uses the same number of cuts as EBOND, and all pre-trained variants use the same molecules and Transformer architecture. Each model is then fine-tuned on the USPTO-50K training set (Schneider et al., 2016), with a model trained without pre-training (None) included as the baseline. Performance is evaluated using Top-k accuracy, round-trip accuracy (Schwaller et al., 2020), and ChemCensor Score (Zagribelnyy et al., 2026). Detailed definitions of the evaluation metrics are provided in Appendix D.

Table 5: **Retrosynthesis results on USPTO-50K with different pretraining strategies.** ChemCensor Max and Top-k denote the maximum and mean over the top-k unique predictions.

Pretrain	**Top-k Accuracy (%) ↑**				**Round-Trip Accuracy (%) ↑**				**ChemCensor Score (CC) ↑**				
	Top-1	Top-3	Top-5	Top-10	RT@1	RT@3	RT@5	RT@10	Max	Top-1	Top-3	Top-5	Top-10
None	48.59	67.07	73.64	81.07	76.11	64.55	57.25	37.73	4.39	3.28	2.17	1.69	1.02
Random	48.03	66.85	74.32	82.58	76.59	66.56	60.70	39.45	4.40	3.30	2.22	1.77	1.07
BRICS	49.11	66.89	73.92	81.61	76.59	67.59	61.70	39.83	4.44	3.29	2.25	1.80	1.08
EBOND	**51.65**	**70.50**	**77.23**	**84.46**	**78.87**	**70.64**	**64.29**	**44.73**	**4.48**	**3.43**	**2.31**	**1.83**	**1.14**
Δ (vs. Best)	***+2.54***	***+3.43***	***+2.91***	***+1.88***	***+2.28***	***+3.05***	***+2.59***	***+4.90***	***+0.04***	***+0.13***	***+0.06***	***+0.03***	***+0.06***

As shown in Table 5, EBOND achieves the best results across all three metrics. Top-1 accuracy increases from 48.59% without pre-training to 51.65%, while Top-10 accuracy increases from 81.07% to 84.46%. Round-trip accuracy also improves across all beam sizes, with RT@10 increasing from 37.73% to 44.73%. EBOND further achieves the highest ChemCensor scores across all beam sizes. These results show that EBOND-derived fragmentation provides useful pre-training signals for retrosynthesis, improving prediction accuracy while supporting more reliable alternative predictions.

Table 6: **Fragment-based molecular optimization on DRD4.** Property and solved rates are computed over valid molecules. Sweet-spot solved rates are computed over valid sweet-spot molecules.

Fragment	Property Optimization ↑				Synthetic Accessibility ↑	
	MW (%)	cLogP (%)	TPSA (%)	Sweet spot (%)	Solved (%)	Sweet-spot solved (%)
BRICS	**82.23**	49.98	68.88	37.31	46.41	33.77
EBOND	75.49	**63.61**	**74.33**	**51.39**	**59.65**	**64.21**
Δ (vs. Best)	-6.74	+13.63	+5.45	+14.08	+13.24	+30.44

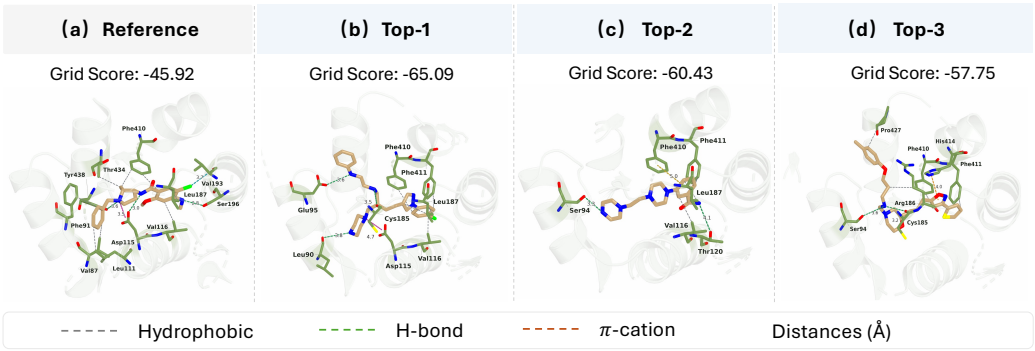


Figure 7: **Representative DRD4 docking results.** (a) Crystal reference ligand Nemonapride in the human dopamine D4 receptor (PDB ID: 5WIV). (b–d) Top-ranked eBOND-generated molecules. Key binding-site interactions and distances are annotated for each pose.

5.2 Fragment-Based Molecular Optimization

We evaluate whether EBOND fragments can serve as generation units for molecular optimization. Following DeepFMPO (Ståhl et al., 2019), lead molecules targeting the dopamine D4 receptor (DRD4) are optimized using either BRICS or EBOND fragments. The optimization model, objectives, and generation budget are kept unchanged, so the two settings differ only in the fragments used for generation. Molecules generated over the final 10 epochs are evaluated under the same budget of 5,120 generation attempts.

We define the sweet spot as $320 \leq \text{MW} \leq 420$, $2.0 \leq \text{cLogP} \leq 3.0$, and $40 \leq \text{TPSA} \leq 60$. As shown in Table 6, the fraction of valid molecules satisfying sweet-spot constraints increases from 37.31% with BRICS to 51.39% with EBOND. The largest gains are observed for cLogP and TPSA. Synthetic accessibility is further evaluated using AiZynthFinder (Genheden et al., 2020). The solved rate increases from 46.41% to 59.65%, while the solved rate among sweet-spot molecules increases from 33.77% to 64.21%. We further conduct a docking-based case study using the human dopamine D4 receptor co-crystallized with Nemonapride (PDB ID: 5WIV) (Wang et al., 2017). Among 200 sampled molecules from each method, 45.0% of EBOND-generated molecules achieve more favorable grid scores than Nemonapride, compared with 39.5% for BRICS. Representative binding modes are shown in Figure 7. These results show that the electronic information captured by EBOND can be translated into useful fragment representations for molecular optimization.

6 Conclusion

We introduced EBOND, a large-scale quantum-chemical dataset for electron-guided molecular fragmentation, covering 452,775 molecules with atom- and bond-level electronic information. Our analysis show that molecule-specific electronic responses provide useful signals for defining fragmentation boundaries and capture synthesis-relevant disconnection patterns beyond rule-based methods. The resulting fragments further expand the rule-based fragment space with distinct physicochemical properties while maintaining strong support from commercially available building blocks. Beyond data analysis, EBOND-derived fragments are useful for retrosynthesis and fragment-based molecular optimization. These results demonstrate the effectiveness of electron-guided fragmentation and establish EBOND as a reusable resource for fragment-based molecular learning.

AI USE STATEMENT

Large language models (LLMs) were used to assist with writing and editing. All AI-assisted content was reviewed and verified by the authors, who take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not raise specific ethical concerns.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of dataset construction, quantum-chemical calculations, electronic descriptors, fragmentation methods, and evaluation settings in the main paper and appendix. The code for dataset construction and quantum-chemical calculations, together with the data and code required to reproduce our experiments, is publicly available at `https://anonymous.4open.science/r/EBOND`.

REFERENCES

Richard FW Bader. Atoms in molecules. *Accounts of chemical research*, 18(1):9–15, 1985.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15 (3):1652–1671, 2019.

Axel D Becke. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of chemical physics*, 98(7):5648–5652, 1993.

M. Born and R. Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389 (20):457–484, 1927. doi: https://doi.org/10.1002/andp.19273892002. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.19273892002`.

Yujie Chen, Tengfei Ma, Zhou Yu, Jiayi Zhang, Ajie Lin, Shu Wu, Dongsheng Cao, and Yiping Liu. Beyond reaction data: Learning chemical knowledge from large-scale molecules for retrosynthesis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 10681–10692, 2026.

Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.

Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using'drug-like'chemical fragment spaces. *ChemMedChem*, 3(10):1503, 2008.

Yanyan Diao, Feng Hu, Zihao Shen, and Honglin Li. Macfrag: segmenting large-scale molecules to obtain diverse fragments with high qualities. *Bioinformatics*, 39(1):btad012, 2023.

Luis R Domingo, Patricia Pérez, and José A Sáez. Understanding the local reactivity in polar organic reactions through electrophilic and nucleophilic parr functions. *RSC advances*, 3(5):1486–1494, 2013.

Daniel A Erlanson, Robert S McDowell, and Tom O'Brien. Fragment-based drug discovery. *Journal of medicinal chemistry*, 47(14):3463–3482, 2004.

M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark,

J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian˜16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

Paul Geerlings, Frank De Proft, and Wilfried Langenaeker. Conceptual density functional theory. *Chemical reviews*, 103(5):1793–1874, 2003.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

M Gonzalez-Suarez, Arie Aizman, Jorge Soto-Delgado, and R Contreras. Bond fukui functions as descriptor of the electron density reorganization in π conjugated systems. *The Journal of Organic Chemistry*, 77(1):90–95, 2012.

Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.

Thomas A Halgren. Mmff vi. mmff94s option for energy minimization studies. *Journal of computational chemistry*, 20(7):720–729, 1999.

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964.

Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273, 2022.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Shao Jinsong, Jia Qifeng, Chen Xing, Yajie Hao, and Li Wang. Molecular fragmentation as a crucial step in the ai-based drug development pathway. *Communications Chemistry*, 7(1):20, 2024.

Benno Kaech, Luis Wyss, Karsten Borgwardt, and Gianvito Grasso. Refine drugs, don't complete them: Uniform-source discrete flows for fragment-based drug discovery. In *International Conference on Learning Representations*, volume 2026, pp. 120938–120963, 2026.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

Greg Landrum, Paolo Tosco, Brian Kelley, Ric, sriniker, gedeck, Riccardo Vianello, David Cosgrove, NadineSchneider, Eisuke Kawashima, Dan N, Andrew Dalke, Gareth Jones, Brian Cole, Matt Swain, Samo Turk, AlexanderSavelyev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Daniel Probst, Kazuya Ujihara, Vincent F. Scalfani, guillaume godin, Axel Pahl, Francois Berenger, JLVarjo, strets123, JP, and DoliathGavid. rdkit/rdkit: 2022_09_3 (q3 2022) release, December 2022. URL `https://doi.org/10.5281/zenodo.7415128`.

Chengteh Lee, Weitao Yang, and Robert G Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical review B*, 37(2):785, 1988.

Seul Lee, Karsten Kreis, Srimukh P Veccham, Meng Liu, Danny Reidenbach, Saee Paliwal, Arash Vahdat, and Weili Nie. Molecule generation with fragment retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:132463–132490, 2024.

Xiao Qing Lewell, Duncan B Judd, Stephen P Watson, and Michael M Hann. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of chemical information and computer sciences*, 38(3):511–522, 1998.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

Tian Lu. A comprehensive electron wavefunction analysis toolbox for chemists, multiwfn. *The Journal of chemical physics*, 161(8), 2024.

Tian Lu and Feiwu Chen. Multiwfn: A multifunctional wavefunction analyzer. *Journal of computational chemistry*, 33(5):580–592, 2012.

Tian Lu and Qinxue Chen. Realization of conceptual density functional theory and information-theoretic approach in multiwfn program. *Conceptual Density Functional Theory: Towards a New Chemical Reactivity Theory*, 2:631–647, 2022.

I Mayer and P Salvador. Overlap populations, bond orders and valences for 'fuzzy'atoms. *Chemical Physics Letters*, 383(3-4):368–375, 2004.

Christophe Morell, André Grand, and Alejandro Toro-Labbé. New dual descriptor for chemical reactivity. *The journal of physical chemistry A*, 109(1):205–212, 2005.

Julianna Oláh, Christian Van Alsenoy, and Abani B Sannigrahi. Condensed fukui functions derived from stockholder charges: assessment of their performance as local reactivity descriptors. *The Journal of Physical Chemistry A*, 106(15):3885–3890, 2002.

Robert G Parr and Weitao Yang. Density functional approach to the frontier-electron theory of chemical reactivity. *Journal of the American Chemical Society*, 106(14):4049–4050, 1984.

Ricardo Pino-Rios. Condensed-to-bond dual descriptor for the analysis of electron density rearrangement in π-conjugated compounds. *Journal of Mathematical Chemistry*, 63(7):1588–1597, 2025.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014.

Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016.

Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12):3316–3325, 2020.

Yu Shee, Anthony M Smaldone, Anton Morgunov, Gregory W Kyro, and Victor S Batista. Fragmentretro: A quadratic retrosynthetic method based on fragmentation algorithms. *Journal of Chemical Theory and Computation*, 2026.

Alexander N Shivanyuk, Sergey V Ryabukhin, A Tolmachev, AV Bogolyubsky, DM Mykytenko, AA Chupryna, W Heilman, and AN Kostyuk. Enamine real database: Making chemical diversity real. *Chemistry today*, 25(6):58–59, 2007.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems*, 34:9405–9415, 2021.

Niclas Ståhl, Goran Falkman, Alexander Karlsson, Gunnar Mathiason, and Jonas Bostrom. Deep reinforcement learning for multiparameter optimization in de novo drug design. *Journal of chemical information and modeling*, 59(7):3166–3176, 2019.

Sheng Wang, Daniel Wacker, Anat Levit, Tao Che, Robin M Betz, John D McCorvy, AJ Venkatakrishnan, Xi-Ping Huang, Ron O Dror, Brian K Shoichet, et al. D4 dopamine receptor high-resolution structures enable the discovery of selective agonists. *Science*, 358(6361):381–386, 2017.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.

Hongxin Xiang, Ke Li, Mingquan Liu, Zhixiang Cheng, Bin Yao, Wenjie Du, Jun Xia, Li Zeng, Xin Jin, et al. Edbench: Large-scale electron density data for molecular modeling. *Advances in Neural Information Processing Systems*, 38, 2026.

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. Mars: Markov molecular sampling for multi-objective drug discovery. *arXiv preprint arXiv:2103.10432*, 2021.

Weijun Xu and Congbao Kang. Fragment-based drug design: from then until now, and toward the future, 2025.

Ruoqi Yang, Hao Zhou, Fan Wang, and Guangfu Yang. Digfrag as a digital fragmentation method used for artificial intelligence-based drug design. *Communications Chemistry*, 7(1):258, 2024.

Weitao Yang and Wilfried J Mortier. The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines. *Journal of the American Chemical Society*, 108(19): 5708–5711, 1986.

Haiyang Yu, Meng Liu, Youzhi Luo, Alex Strasser, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Qh9: A quantum hamiltonian prediction benchmark for qm9 molecules. *Advances in Neural Information Processing Systems*, 36:40487–40503, 2023.

Yuliana Zabolotna, Dmitriy M Volochnyuk, Sergey V Ryabukhin, Kostiantyn Gavrylenko, Dragos Horvath, Olga Klimchuk, Oleksandr Oksiuta, Gilles Marcou, and Alexandre Varnek. Synthi: a new open-source tool for synthon-based library design. *Journal of Chemical Information and Modeling*, 62(9):2151–2163, 2021.

Bogdan Zagribelnyy, Ivan Ilin, Maksim Kuznetsov, Nikita Bondarev, Roman Schutski, Thomas MacDougall, Rim Shayakhmetov, Zulfat Miftahutdinov, Mikolaj Mizera, Vladimir Aladinskiy, Alex Aliper, and Alex Zhavoronkov. When single answer is not enough: Rethinking single-step retrosynthesis benchmarks for LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ABhGgV7pov.

Leili Zhang, Vasumitra Rao, and Wendy Cornell. r-brics–a revised brics module that breaks ring structures and carbon chains. *ChemMedChem*, 19(4):e202300202, 2024.

Shuai Zhang, Junfeng Fang, Xuqiang Li, Hongxin Xiang, Jun Xia, Ye Wei, Wenjie Du, and Yang Wang. Iterative substructure extraction for molecular relational learning with interactive graph information bottleneck. In *International Conference on Learning Representations*, volume 2025, pp. 63678–63706, 2025.

APPENDIX

A QUANTUM-CHEMICAL BACKGROUND

A.1 ELECTRONIC STRUCTURE AND DENSITY FUNCTIONAL THEORY

Under the Born–Oppenheimer approximation, nuclear and electronic motions are approximately separated, allowing the electronic structure to be evaluated at a fixed nuclear geometry (Born & Oppenheimer, 1927). For fixed nuclear coordinates, the electronic state is formally described by the many-electron Schrödinger equation, $\hat{H}_e\Psi = E\Psi$, where $\hat{H}_e$ is the electronic Hamiltonian, Ψ is the many-electron wave function, and E is the corresponding electronic energy. Direct treatment of the many-electron wave function becomes rapidly more expensive as the number of electrons increases, since Ψ depends on the coordinates of all electrons simultaneously.

Density functional theory (DFT) provides a practical alternative by using the electron density $\rho(\mathbf{r})$ as its basic variable rather than the full many-electron wave function. The first Hohenberg–Kohn theorem establishes that the ground-state electron density uniquely determines the external potential, up to an additive constant, and therefore determines the ground-state properties of the system (Hohenberg & Kohn, 1964). This result provides the theoretical basis for describing an interacting electronic system through its electron density.

Kohn–Sham DFT makes this formulation computationally practical by introducing an auxiliary non-interacting electronic system that reproduces the ground-state density of the interacting system (Kohn & Sham, 1965). The Kohn–Sham equations in atomic units are written as follows:

$$\left[-\frac{1}{2}\nabla^2 + v_{\text{eff}}(\mathbf{r})\right]\phi_i(\mathbf{r}) = \epsilon_i\phi_i(\mathbf{r}),$$

where $\phi_i(\mathbf{r})$ and ϵ_i are the Kohn–Sham orbitals and orbital energies, respectively. The effective potential is defined as $v_{\text{eff}}(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_{\text{H}}(\mathbf{r}) + v_{\text{xc}}(\mathbf{r})$, where v_{ext} describes the interaction between electrons and nuclei, v_{H} is the classical Coulomb contribution from the electron density, and v_{xc} is the exchange–correlation potential that accounts for the remaining many-electron effects.

The electron density is obtained from the occupied Kohn–Sham orbitals as $\rho(\mathbf{r}) = \sum_i n_i|\phi_i(\mathbf{r})|^2$, where n_i denotes the occupation of orbital i. For spin-unrestricted calculations, the total density can be written as $\rho(\mathbf{r}) = \rho_\alpha(\mathbf{r}) + \rho_\beta(\mathbf{r})$, where $\rho_\alpha(\mathbf{r}) = \sum_i n_i^\alpha|\phi_i^\alpha(\mathbf{r})|^2$ and $\rho_\beta(\mathbf{r}) = \sum_i n_i^\beta|\phi_i^\beta(\mathbf{r})|^2$. This spin-resolved representation is particularly useful for open-shell systems.

A.2 Response to Changes in Electron Number

Conceptual density functional theory describes how the electron density changes with the number of electrons at a fixed external potential (Parr & Yang, 1984; Geerlings et al., 2003). The local response is described by the Fukui function, $f(\mathbf{r}) = (\partial\rho(\mathbf{r})/\partial N)_{v(\mathbf{r})}$. At an integer electron number, the responses to electron addition and removal correspond to the right- and left-sided derivatives, respectively. In practice, they are commonly approximated by finite differences, $f^+(\mathbf{r}) \approx \rho_{N+1}(\mathbf{r}) - \rho_N(\mathbf{r})$ and $f^-(\mathbf{r}) \approx \rho_N(\mathbf{r}) - \rho_{N-1}(\mathbf{r})$. Their difference defines the dual descriptor, $\Delta f(\mathbf{r}) = f^+(\mathbf{r}) - f^-(\mathbf{r})$, which measures the difference between the local responses to electron addition and removal (Morell et al., 2005). In EBOND, the N, $N+1$, and $N-1$ states are evaluated at the same nuclear geometry, keeping the external potential fixed so that the finite differences reflect electronic response without contributions from structural relaxation.

A.3 Atom- and Bond-Level Electronic Descriptors

Electronic structure is defined in continuous real space, but local chemical information can be represented at the atom and bond levels. Atomic population analysis assigns electronic descriptors to individual atoms, including atomic charges, spin populations, and condensed response functions (Yang & Mortier, 1986; Domingo et al., 2013). Bond-level analysis describes electronic interactions between pairs of atoms through bond-order descriptors and their response to changes in electron number (Mayer & Salvador, 2004; Gonzalez-Suarez et al., 2012). These local representations connect continuous electronic structure with the atom- and bond-level descriptors used in EBOND. The specific descriptors and their definitions are given in Section 3.

B Dataset Construction Details

Molecule sampling. The molecular collection was constructed from a diverse set of product molecules sampled from USPTO-FULL together with molecules involved in USPTO-50K reactions. Starting from 911,610 USPTO products, invalid structures, non-carbon molecules, formally charged molecules, and molecules containing metallic elements or any of Ge, As, Sb, and Te were removed. After canonicalization, deduplication, and exclusion of previously processed molecules, 864,803 candidates remained. These candidates were grouped by heavy-atom count (HAC), with sampling quotas assigned according to the size distribution of the filtered pool. Within each HAC group, structural redundancy was reduced by limiting the number of molecules sharing the same Bemis–Murcko scaffold, followed by Max-Min diversity sampling using radius-2, 2,048-bit Morgan fingerprints. This procedure selected 450,000 product molecules from USPTO-FULL. Molecules involved in USPTO-50K reactions were then added to preserve the benchmark reaction space. After the two sources were combined and deduplicated, 538,998 unique molecules were retained for electronic-structure calculations.

Quality control. We performed quality control throughout the electronic-structure calculation pipeline. A valid MMFF94s conformer was required before geometry optimization, and optimized

structures were retained only after passing the imaginary-frequency check. For each retained structure, the DFT calculations for the N, $N+1$, and $N-1$ states were required to reach SCF convergence. Finally, all required electronic descriptors were successfully extracted with Multiwfn. Table 7 summarizes the calculation and quality-control statistics, with 452,775 of the 538,998 submitted molecules passing the full pipeline.

Table 7: **Quality-control statistics.** Success is relative to the previous stage.

Stage	Molecules (#)	Success (%)
Submitted for calculation	538,998	100.00
MMFF94s conformer generation	534,544	99.17
GFN2-xTB optimization and frequency check	457,813	85.65
Three-state DFT calculation	452,777	98.90
Multiwfn descriptor extraction	452,775	100.00
Overall success	452,775	84.00

Computational resources. All calculations were performed on the high-performance computing cluster described in Section 3.1. MMFF94s conformer generation, GFN2-xTB optimization, three-state DFT calculations, and Multiwfn analysis used 4, 1–2, 16, and 16 CPU cores per molecule, respectively, with 48 GB of memory allocated to each DFT calculation. The corresponding time limits were 0.5, 0.5, 5, and 1 hour per molecule. Overall, constructing EBOND required approximately 1.26 million CPU core-hours, equivalent to 144.2 years of serial computation on a single CPU core, with three-state DFT calculations accounting for 91.5% of the total computational cost.

C Dataset Details

C.1 Dataset Format and Fields

The released EBOND dataset provides molecular structures together with atom- and bond-level electronic descriptors and the corresponding fragmentation scores. Each molecule is assigned a unique identifier. The main fields are summarized in Table 8.

Table 8: **Main fields in the EBOND dataset.**

Level	Field	Description
Molecule	Molecule ID	Unique molecule identifier
Molecule	SMILES	Molecular structure representation
Atom	Atom index	Atom index within the molecule
Atom	Element	Atomic element
Atom	CDD	Condensed dual descriptor
Atom	P^+, P^-	Electrophilic and nucleophilic Parr functions
Bond	Atom pair	Indices of the bonded atoms
Bond	Bond type	Bond type in the molecular graph
Bond	BDD	Bond dual descriptor
Fragmentation	$\pi_{\mathrm{BDD}}^{\mathrm{elec}}, \pi_{\mathrm{BDD}}^{\mathrm{nuc}}$	BDD-based fragmentation scores
Fragmentation	π_{CDD}	CDD-based fragmentation score
Fragmentation	π_{Parr}	Parr-based fragmentation score
Fragmentation	Selected bond	Final fragmentation boundary selected by EBOND
Fragmentation	Fragment	fragments generated from the selected bonds

C.2 Molecular Statistics

The EBOND dataset contains 452,775 unique molecules and 4,709,284 non-ring heavy-atom candidate bonds. The molecules contain 10 non-hydrogen elements and cover 67 heavy-atom bond types.

In addition to the dataset-level statistics reported in Figure 3, we summarize several molecular properties to further describe the chemical space covered by EBOND. As shown in Table 9, the dataset spans broad ranges of molecular size, lipophilicity, polarity, flexibility, and carbon hybridization.

Table 9: **Molecular properties of the EBOND.** P5 and P95 denote the 5th and 95th percentiles.

Property	Mean	Median	P5	P95
MW (Da)	357.35	346.18	172.29	579.65
cLogP	3.43	3.32	0.55	6.68
TPSA (Å^2)	67.77	64.15	20.23	127.52
RotB	4.73	4.00	1.00	11.00
Fsp3	0.36	0.33	0.00	0.83

C.3 Electronic Descriptor Statistics

We further provide the distributions of the electronic descriptors provided in EBOND. Figure 8 shows the distributions of CDD, P^+, and P^- across the ten non-hydrogen elements. Figure 9 shows BDD distributions across the twelve most frequent non-ring heavy-atom bond types.

C.4 Dataset Scope and Limitations

EBOND is constructed from USPTO-derived molecules to focus on molecules represented in recorded synthetic reactions. USPTO is widely used in data-driven reaction prediction and retrosynthesis, making it a natural source for studying fragmentation in a synthesis-relevant chemical space. The current release covers carbon-containing organic molecules with elements H, C, N, O, F, P, S, Cl, Br, I, and Si. Molecules containing metals, as well as Ge, As, Sb, and Te, were excluded to maintain the consistency of large-scale quantum-chemical calculations. The current fragmentation setting focuses on non-ring heavy-atom bonds. Future extensions can broaden EBOND to more diverse chemical spaces and fragmentation settings.

D Evaluation Metrics

D.1 Data Analysis Metrics

We use $\mathcal{D}_{\mathrm{rxn}}$ to denote the reaction evaluation set, $\mathcal{B}(M)$ to denote the non-ring heavy-atom bonds considered for fragmentation, and $\mathcal{B}_m(M)$ to denote the bonds selected by method m.

Recorded-disconnection coverage. Recorded-disconnection coverage (RDC) measures the fraction of recorded disconnections selected by fragmentation method m:

$$\mathrm{RDC}_m = \frac{|\{(M,b) \in \mathcal{D}_{\mathrm{rxn}} \mid b \in \mathcal{B}_m(M)\}|}{|\mathcal{D}_{\mathrm{rxn}}|}.$$

To measure the coverage after combining EBOND with a rule-based method r, we define combined recorded-disconnection coverage (C-RDC) as follows:

$$\text{C-RDC}_r = \frac{|\{(M,b) \in \mathcal{D}_{\mathrm{rxn}} \mid b \in \mathcal{B}_r(M) \cup \mathcal{B}_{\mathrm{eBOND}}(M)\}|}{|\mathcal{D}_{\mathrm{rxn}}|}.$$

Here, r may denote either a single rule-based method or a union of rule-based methods.

Missed-disconnection recovery. For a rule-based reference method r, we define the recorded disconnections within the fragmentation scope that are missed by r as $\mathcal{D}^r_{\mathrm{miss}} = \{(M,b) \in \mathcal{D}_{\mathrm{rxn}} \mid b \in \mathcal{B}(M),\, b \notin \mathcal{B}_r(M)\}$. Missed-disconnection recovery (MDR) measures the fraction of these missed disconnections that are selected by EBOND:

$$\mathrm{MDR}_r = \frac{|\{(M,b) \in \mathcal{D}^r_{\mathrm{miss}} \mid b \in \mathcal{B}_{\mathrm{eBOND}}(M)\}|}{|\mathcal{D}^r_{\mathrm{miss}}|}.$$

Building-block coverage. For fragmentation method m, $\mathcal{V}_m$ denotes the attachment-aware fragment vocabulary, where the original connection sites are retained. The standardized Enamine US Stock

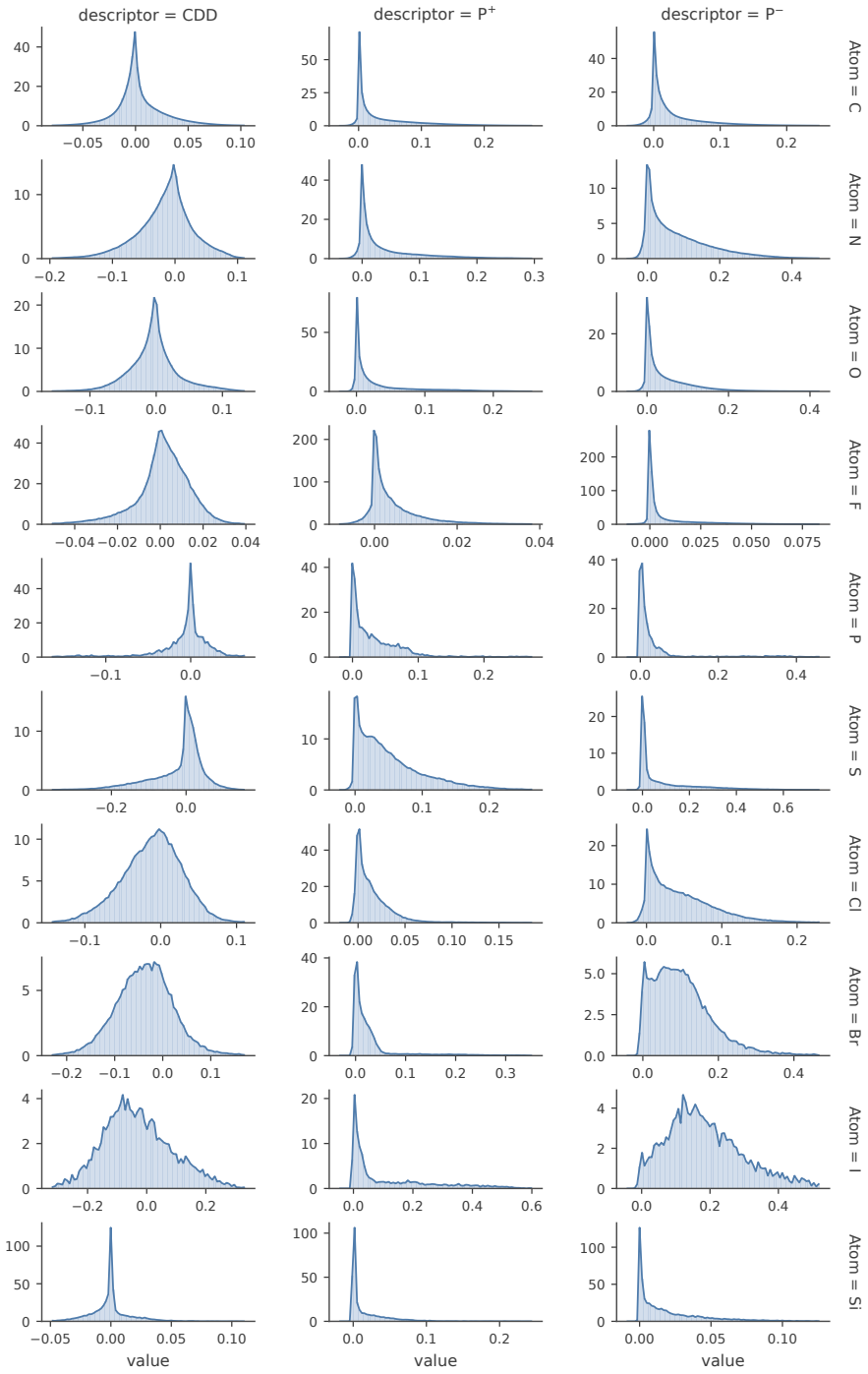


Figure 8: **Distributions of atom-level electronic descriptors in EBOND.** Rows correspond to the ten non-hydrogen elements, and columns show CDD, P^+, and P^-, respectively. Each distribution is normalized independently for visualization.

collection is denoted by $\mathcal{E}$. A fragment u is considered covered if it matches a substructure of at least one compound $e \in \mathcal{E}$, denoted by $u \preceq e$. Building-block coverage (BCR) is defined as follows:

$$\mathrm{BCR}_m = \frac{|\{u \in \mathcal{V}_m \mid \exists e \in \mathcal{E},\, u \preceq e\}|}{|\mathcal{V}_m|}.$$

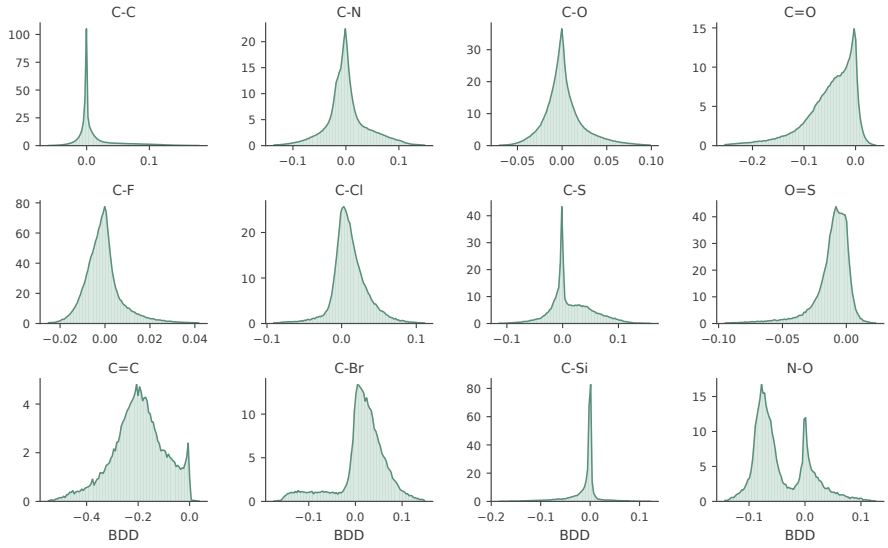


Figure 9: **Distributions of bond-level electronic descriptors in EBOND.** BDD distributions are shown for the twelve most frequent non-ring heavy-atom bond types considered for fragmentation. Each distribution is normalized independently for visualization.

D.2 Downstream Application Metrics

D.2.1 Retrosynthesis

We evaluate retrosynthesis using Top-k Accuracy, Round-Trip Accuracy, and ChemCensor metrics. For the i-th test example, P_i denotes the target product, Y_i^{true} the ground-truth reactant set, and $\hat{Y}_i^j$ the j-th predicted reactant set. The corresponding predicted retrosynthetic reaction is denoted by $R_i^j = (P_i, \hat{Y}_i^j)$.

Top-k Accuracy. Top-k Accuracy measures whether the recorded reactant set is recovered among the top-k predictions. For a test set of N products, it is defined as follows:

$$\text{Top-}k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[Y_i^{\text{true}} \in \{\hat{Y}_i^j \mid j = 1, \ldots, k\}\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Round-Trip Accuracy. Round-Trip Accuracy evaluates whether predicted reactants can reproduce the target product using a forward reaction model (Schwaller et al., 2020). For the top-k predictions, it is defined as follows:

$$\text{RT@}k = \frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k} \mathbb{I}\left[F(\hat{Y}_i^j) = P_i\right],$$

where $F(\cdot)$ denotes the forward reaction model.

ChemCensor. ChemCensor (CC) evaluates the plausibility of a predicted reaction using reaction-center and functional-group precedents from recorded reactions (Zagribelnyy et al., 2026). Each predicted reaction receives an integer score from 0 to 5, with larger values indicating stronger support from recorded synthetic precedents. For product i, $\mathcal{R}_i$ denotes the set of predicted retrosynthetic reactions evaluated by ChemCensor. The maximum ChemCensor score is defined as follows:

$$\text{MaxCC}(i) = \max_{R \in \mathcal{R}_i} \text{CC}(R).$$

The average per-target maximum ChemCensor score is then computed as follows:

$$\text{Av. PT-Max CC} = \frac{1}{N}\sum_{i=1}^{N} \text{MaxCC}(i).$$

We further report the average ChemCensor score over the top-k unique predicted reactions. For product i, $\widetilde{\mathcal{R}}_i^{(k)}$ denotes the set of unique reactions among $\{R_i^j\}_{j=1}^k$. The score is defined as follows:

$$\mathrm{CC@}k(i) = \frac{1}{k} \sum_{R \in \widetilde{\mathcal{R}}_i^{(k)}} \mathrm{CC}(R).$$

The average per-target Top-k ChemCensor score is then defined as follows:

$$\text{Av. PT-Top-k CC} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CC@}k(i).$$

Following the ChemCensor setting, duplicate predictions are removed, and missing positions are assigned a score of zero when fewer than k unique predictions remain.

D.2.2 Fragment-based Molecular Optimization

We evaluate fragment-based molecular optimization using Sweet-Spot Rate and Solved Rate. $\mathcal{X}$ denote the set of molecules to be evaluated.

Sweet-Spot Rate. A molecule is considered a sweet-spot molecule if all target properties fall within their predefined ranges (Ståhl et al., 2019). Sweet-Spot Rate measures the fraction of evaluated molecules that satisfy all optimization objectives and is defined as follows:

$$\text{Sweet-Spot Rate} = \frac{\sum_{x \in \mathcal{X}} \mathbb{I}[S(x) = 1]}{|\mathcal{X}|},$$

where $S(x) = 1$ indicates that molecule x satisfies all predefined target-property ranges.

Solved Rate. Solved Rate evaluates molecular synthesizability using a retrosynthesis planner. A molecule is considered solved if the planner identifies a complete synthetic route to available starting materials. It is defined as follows:

$$\text{Solved Rate} = \frac{\sum_{x \in \mathcal{X}} \mathbb{I}[\mathrm{Solve}(x) = 1]}{|\mathcal{X}|},$$

where $\mathrm{Solve}(x) = 1$ indicates that a complete synthetic route is found for molecule x.

E Additional Experimental Results

E.1 Contribution of Electronic Descriptors

To examine the contribution of the three electronic descriptors, we evaluate BDD, CDD, and Parr functions individually and perform remove-one ablations under the same fragmentation setting used in the main experiments. As shown in Table 10, the full combination achieves the highest RDC, with BDD contributing the largest gain among the three descriptors.

Table 10: Contribution of electronic descriptors to disconnection recovery.

Setting	Avg. Cuts/mol	RDC (%)
BDD	5.09	68.86
CDD	1.54	35.82
Parr	3.83	61.62
w/o BDD	4.09	64.38
w/o CDD	5.98	79.73
w/o Parr	5.59	76.23
All	6.17	**81.10**

E.2 Selection of the Fragmentation Threshold

The threshold τ controls the balance between disconnection coverage and the number of selected bonds. As shown in Table 11, $\tau = 0.25$ gives high disconnection coverage without selecting too many bonds on the USPTO-50K validation split. The selected threshold was then fixed and applied to the full USPTO-50K evaluation set used in the main experiments.

Table 11: Fragmentation threshold selection on the USPTO-50K validation split.

τ	Avg. Cuts / mol	RDC (%)	MDR (%)
-0.50	4.09	65.09	60.67
-0.25	4.55	69.86	64.50
0.00	5.15	73.62	67.33
0.25	**6.10**	**80.77**	**73.67**
0.50	9.09	91.04	83.00