Exact Total Sigma-Bond Counting from Composition via a Mass-Derived Descriptor and B-Matching

Luke Miller*

Independent Researcher, United States

E-mail: lukemiller@tutanota.com

Abstract

We present a geometry-free algorithm that computes the exact total number of σ bonds in an organic molecule from elemental composition alone. Each atom receives a single mass-derived rung descriptor; pair propensities are scored by a fixed dual-peak kernel; bonds are assigned by an exact weighted b-matching solved via min-cost max-flow. Within scope (neutral formulas over {H,C,N,O,F,Cl,Br,I,P,S}), the method is deterministic and parameter-light. We first verified exact agreement on a curated set from PubChem (673 valid formulas after excluding malformed or out-of-scope entries). We then performed a large-scale validation on *ChEMBL 36*, sampling 100,000 formulas that satisfy the same scope criteria. The algorithm matched the valence identity $\frac{1}{2}\sum_i \text{valence}(i)$ for all 100,000 cases (100.0%), without 3D coordinates or training. These results support the use of the method as a robust primitive for QSPR/QSAR pipelines.

Introduction

Counting bonds underpins cheminformatics, property prediction, and early-stage screening. Conventional routes either rely on hand-coded valence rules or invoke quantum chemistry with geometry optimization. We offer a third path. A single global descriptor derived from atomic mass, together with a universal resonance kernel and an exact combinatorial optimizer, yields the total number of sigma bonds directly from a molecular formula. The approach is deterministic and requires neither learned parameters nor 3D coordinates. It complements descriptor-based QSPR/QSAR methods by supplying a robust structural primitive.[1,2] We subsequently validated exactness at scale on a 100,000-molecule sample from ChEMBL 36.

Conceptual contribution. The element–level descriptor we use is not a chemoinformatic feature tuned on molecules. It is imported unchanged from a rigid, two-anchor logarithmic mass ladder fixed by the electron and muon (m_e, m_μ) in a closed spectral construction.[3] This yields a single scalar "rung" ν_Z per element with no parameters learned from chemistry. Together with one universal dual-peak kernel on $|\nu_i - \nu_j|$ and an exact weighted b-matching, this gives a geometry-free, training-free rule that allocates bonds from composition alone. The novelty is the *physics-anchored allocation rule*: from only a molecular formula, the method deterministically selects which pairs are favored to bond by maximizing a mass-derived resonance score under valence capacity, without any fit to molecular datasets.

Scope and constants

We treat neutral organic molecules using common valence capacities (H 1, C 4, N 3, O 2, F 1, Cl 1, Br 1, I 1, P 3, S 2). Atomic masses are standard values; isotopic specificity is not required. Lepton anchors are PDG 2024.[4] Only the electron and muon define the rung map and its step $g = (m_\mu/m_\mathrm{e})^{1/13} = 1.507003107$. This is the same locked constant used in a companion manuscript on masses. The algorithm reports the total number of sigma bonds, not connectivity or pi systems. All validations reported here enforce this scope at dataset-construction time.

Method

The algorithm proceeds in three stages: (1) assign each atom a single scalar descriptor derived from its mass, (2) score the resonance affinity for all atom pairs using a universal kernel, and (3) solve for the bond count using an exact combinatorial optimizer.

The Mass-Derived Rung Descriptor

The core of the algorithm is a single scalar descriptor, the "rung index" ν_Z, assigned to each element Z. This descriptor is not an empirical feature tuned on chemical data, but is imported without modification from a universal spectral ladder developed in a companion work that organizes the mass spectrum of fundamental particles.[3]

The ladder is a logarithmic ruler anchored by the electron and muon masses. By fixing the electron at rung $k = 3$ and the muon at $k = 16$, the mapping from an atomic mass m to its rung $k(m)$ is uniquely defined as:

$$k(m) = 3 + \frac{13\ \ln(m/m_\mathrm{e})}{\ln(m_\mu/m_\mathrm{e})} \tag{1}$$

The step-factor is locked to the public PDG constant, $g = (m_\mu/m_e)^{1/13} = 1.507003107$. The constants in Eq. (1) are therefore fixed by fundamental physics, not by this chemical model. Applying this equation to neutral atomic masses yields the final descriptor ν_Z for each element, shown for a representative set in Table 1. These values are the only per-element inputs to the bond-counting algorithm.

Table 1: Representative element rungs $\nu_Z = k(m_Z)$ with $g = 1.507003107$. Values are dimensionless.

Z	H	C	N	O	P	S
ν_Z	21.326	27.368	27.743	28.067	29.679	29.756

Universal Resonance Kernel

Given a pair of atoms (i, j) with rung indices (ν_i, ν_j), their resonance affinity s_{ij} is calculated using a fixed, universal dual-peak kernel based on their rung separation $\Delta\nu = |\nu_i - \nu_j|$:

$$s_{ij} = 1.124462 \exp\Big(-(\Delta\nu/0.90)^2\Big) + 1.551250 \exp\Big(-\frac{(\Delta\nu - 6.50)^2}{2\,1.20^2}\Big) \tag{2}$$

The first term favors like-like pairs ($\Delta\nu \approx 0$), while the second favors specific heavy-light contrasts ($\Delta\nu \approx 6.50$). The kernel parameters $(1.124462, 1.551250, 0.90, 1.20, 6.50)$ are universal constants, fixed for all molecules.

Exact Bond Assignment via b-Matching

The final step is to assign bonds to maximize the total resonance affinity $\sum s_{ij}$, subject to the valence capacity b_i of each atom. This is a classic weighted b-matching problem on a complete graph of the molecule's atoms. We solve this problem to optimality by reformulating it as a min-cost max-flow problem on a bipartite network.[5,6] Because all capacities and supplies are integers, the solution is guaranteed to be integral, yielding an exact integer count for the total number of sigma bonds. The implementation uses Johnson's algorithm for all-pairs shortest paths with deterministic tie-breaking to ensure reproducibility.

Validation and results

We report two checks with distinct purposes. First, a *scope sanity check*: for neutral formulas over {H,C,N,O,F,Cl,Br,I,P,S}, the valence identity enforces $\frac{1}{2}\sum_i \text{valence}(i) = |E_\sigma|$. Our min-cost max-flow construction on a complete graph saturates these degrees by design; consequently, the predicted total equals the identity on the PubChem baseline (673/673) and on a 100,000-formula sample from ChEMBL 36 (100,000/100,000). This establishes feasibility, determinism, and scale readiness within scope.

Second, to test generalization at scale, we constructed a large validation set from *ChEMBL 36*. We used the official SQLite dump and joined the small-molecule registry to the compound-properties table to obtain names and molecular formulas. We filtered to neutral formulas that consist only of the symbols {H,C,N,O,F,Cl,Br,I,P,S} and that do not contain charge/adduct markers or other annotations (such as plus or minus signs, parentheses, dots, or whitespace). From the formulas that passed these criteria, we drew a random sample of 100,000 for evaluation. On this dataset the algorithm matched the valence identity for every case (100.0%), again without geometry or learned parameters.

The substantive claim in this work is not discovery of the total $|E_\sigma|$ (which follows from the valence identity inside scope), but the *allocation rule* induced by a physics-anchored descriptor and a universal kernel: given only composition, the method selects a high-affinity simple graph by solving an exact weighted b-matching with fixed constants. The 100,000-molecule result shows that this rule operates stably at scale without geometry or training; its value is that the same closed-form ingredients can be used to propose chemically plausible adjacencies and σ/π partitioning from composition alone.

Table 2: Illustrative examples. The method returns the exact total of sigma bonds using only the molecular formula.

Molecule	Formula	Expected $\Sigma\sigma$	Predicted $\Sigma\sigma$
Methane	CH_4	4	4
Benzene	C_6H_6	12	12
Aspirin	$C_9H_8O_4$	21	21
Ibuprofen	$C_{13}H_{18}O_2$	33	33
Cholesterol	$C_{27}H_{46}O$	80	80
Uridine diphosphate glucose	$C_{15}H_{24}N_2O_{17}P_2$	94	94

Discussion

The rung descriptor yields a single scalar per element that transfers across molecules. The universal kernel encodes generic like-like and heavy–light bias. The optimization is exact and integral by construction, so the reported total is the resonance-saturating count under valence constraints. The combination explains the observed 100% exact agreement on both the PubChem baseline and the 100,000-molecule ChEMBL sample without geometry or training. The method is a practical primitive for large-scale enumeration, retrosynthesis heuristics, and as a stable feature for QSPR/QSAR models. In short, the contribution is a closed-form, physics-anchored rule that maps composition → per-element rung → universal pair score → exact integral allocation, yielding a deterministic bond proposal without geometry or learned parameters. The success of this approach opens several avenues for future research. While the current work focuses on the total sigma-bond count, the underlying resonance scores (s_{ij}) contain rich information about pairwise bonding propensity. Future work could leverage these scores to develop algorithms for full connectivity graph prediction. Furthermore, investigating the physical meaning of the ”like-like” and ”heavy-light” resonance peaks in the universal kernel may yield new insights into the fundamental principles governing chemical bonding.

Scope, assumptions, and limits

The algorithm returns the number of sigma bonds, not the connectivity graph or pi-bond structure. It assumes neutral molecules and standard valence capacities. Hypervalent species and charged states are out of scope in the present form. Extending to additional elements is straightforward once valences are specified. Recovering connectivity and multiplicities would require an augmented objective or a postprocessing stage.

Claims and non-claims. We claim a physics-anchored, geometry-free allocation rule for bonds from composition: a single mass-derived rung per element, a fixed dual-peak kernel on rung separation, and an exact integral optimizer. We do not claim that the total number of σ bonds exceeds the valence identity; inside scope that total is constrained and serves here only as a completeness check at scale. We also do not claim reconstruction of full connectivity or stereochemistry; the present focus is the allocation rule and its deterministic behavior under fixed constants.

Data sources and construction

PubChem was used for the small baseline set and is documented in its bulk downloads and update notes.[7,8] For the large-scale test we used *ChEMBL 36*.[9,10] Names were taken from the preferred name when available (otherwise the registry identifier), and formulas from the *compound properties* schema field that records the full molecular formula.[11,12] The scope filter retained only neutral formulas over {H,C,N,O,F,Cl,Br,I,P,S} and removed records containing charge or adduct notation and other non-formula characters. A two-column sheet (name, formula) was produced for evaluation; no additional properties were used.

Reproducibility

All constants used by the descriptor and kernel are fixed in code. The evaluation requires only a two-column sheet (name, formula). We publish the script that obtains formulas from the public ChEMBL dump, applies the scope filter, and writes the evaluation sheet, together with the exact 100,000-row CSV. Running the field test over this sheet deterministically reproduces the counts and the 100.0% exact-match result.

Data availability

The two-column evaluation sheet derived from *ChEMBL 36* (100,000 rows), the baseline PubChem list (with excluded entries annotated). These datasets can be regenerated directly from the cited public sources following the documented filter.

Code availability

All code to compute rung indices, evaluate the kernel, and run min-cost max-flow is provided as `molecule_fieldtest.py` with a minimal helper. The flow solver follows successive shortest path with potentials.[5,6] The rung, kernel, and valence capacities are the exact constants shown in the code.

Related methods

This work complements graph-based bond-order estimators and population analyses such as DDEC6,[13] and constraint strategies used in MD such as ILVES.[14] It differs in being geometry-free, training-free, and focused on the sigma-bond total.

Acknowledgments

We thank colleagues for feedback on combinatorial optimization and descriptor design. Any remaining errors are the authors' own.

Integrality and optimality

With integer supplies b_i and finite capacities, the min-cost max-flow solution on a bipartite network is integral, so it corresponds to an integer b-matching. This follows from total unimodularity of the node–arc incidence matrix or standard integrality theorems.[5,6] The global optimum maximizes summed resonance under the degree constraints.

References

(1) Cherkasov, A. et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry* **2014**, *57*, 4977–5010.

(2) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29*, 476–488.

(3) Miller, L. A Single Logarithmic Function for the Lepton, Hadron, and Atomic Mass Spectra. Manuscript / preprint, 2025; Companion paper defining the rung map $\nu(m) = 3 + \frac{13\ln(m/m_e)}{\ln(m_\mu/m_e)}$ used here.

(4) Group, P. D. Review of Particle Physics. `https://pdg.lbl.gov/`, 2024; Accessed 2025-10-15.

(5) Ahuja, R. K.; Magnanti, T. L.; Orlin, J. B. *Network Flows: Theory, Algorithms, and Applications*; Prentice Hall, 1993.

(6) Schrijver, A. *Combinatorial Optimization: Polyhedra and Efficiency*; Springer, 2003.

(7) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Research* **2022**, *51*, D1373–D1380.

(8) PubChem Downloads and Bulk Access. `https://pubchem.ncbi.nlm.nih.gov/docs/downloads`, Accessed 2025-10-19.

(9) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; others The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* **2024**, *52*, D1180–D1188.

(10) ChEMBL 36 is live. `https://www.ebi.ac.uk/about/news/updates-from-data-resources/chembl-36/`, EMBL-EBI news announcement, 2025-10-15.

(11) ChEMBL Interface Documentation: Downloads. `https://chembl.gitbook.io/chembl-interface-documentation/downloads`, Current release and download formats.

(12) ChEMBL Schema: compound properties (full_molformula). `https://www.ebi.ac.uk/chembl/api/data/molecule/schema`, Schema field documenting the full molecular formula.

(13) Manz, T. A.; Limas, N. G. Introducing DDEC6 atomic population analysis: part 3. Comprehensive method to compute bond orders. *RSC Advances* **2017**, *7*, 45552–45581.

(14) Campaña, C.; others ILVES: Accurate and Efficient Bond Length and Angle Constraints in Molecular Dynamics. *J. Chem. Theory Comput.* **2018**, *14*, 4296–4308.