PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2026, **28**, 13021

Machine learning the quantum topology of chemical bonds

Michal Michalski *[a] and Slawomir Berski *[b]

Chemical bonding can be characterized within quantum chemical topology (QCT), which provides a real-space description *via* the topological analysis of the electron density and the electron localization function (ELF). While QCT has traditionally been applied on a molecule-by-molecule basis, recent advances in machine learning (ML) and the availability of large quantum chemical datasets now enable bonding analysis at scale. Here, we integrate ELF-based topological descriptors with ML to establish a data-driven framework for mapping chemical bonding across the QM9 dataset. Wavefunctions computed at the B3LYP/6-31G(2df,p) level were used to extract ELF basin populations, which were paired with geometric and bonding descriptors to construct a bond-level dataset. Statistical analysis revealed relationships between ELF populations, bond lengths, and local chemical environments. Regression models were trained to predict ELF electron populations directly from molecular geometry. The best performance was obtained when local environmental descriptors were included, reducing the prediction error by a factor of two relative to models using only the bond type and bond length. These results demonstrate that real-space bonding parameters, such as bond electron populations, can be predicted from simple structural features, enabling scalable and interpretable exploration of chemical bonding across large chemical spaces.

Received 5th January 2026,
Accepted 28th April 2026

DOI: 10.1039/d6cp00029k

rsc.li/pccp

1 Introduction

The concept of the chemical bond has long constituted a fundamental paradigm for rationalizing molecular structures and chemical reactivity.[1,2] Although originally explained by molecular orbital theories,[3] modern approaches based on quantum chemical topology (QCT)[4] have provided a real-space definition of bonding.[5] The atoms in molecules (AIM) theory of Bader[6] defines atomic basins through attractors and gradient paths of an electron density field, offering a partitioning of molecular space. The bonding between atoms is characterized by a critical point of index 1 lying on a gradient path (bond path) linking two attractors.[7] In parallel, the ELF, introduced by Becke and Edgecombe[8] and later refined by Savin and Silvi,[9] quantifies the likelihood of finding paired electrons with parallel spins[10] and serves as a real-space indicator of electron pair localization.[11] Within this framework, a bond corresponds to a region of space characterized by high electron localization, identifiable through the topology of the electron localization function (ELF)[12] or the electron localizability indicator (ELI).[13] From a topological standpoint, the ELF field partitions molecular space into basins separated by zero-flux surfaces, each associated with a local maximum.[14] Two principal basin types are distinguished: monosynaptic basins V(X), which describe lone pairs or non-bonding regions, and disynaptic basins V(A,B), representing shared-electron interactions between atoms A and B.[15] Integration of the electron density within these basins yields populations that quantify electron sharing and covalency, providing an intuitive link between the chemical structure and bonding.[16] The bonding evolution theory (BET)[17] extends this framework by analysing how the ELF topology evolves with molecular geometry along the reaction path. Topological catastrophes correspond to the formation, rearrangement, or disappearance of ELF basins and thus describe the sequence of electron density reorganizations underlying structural changes.[18] This topological perspective translates qualitative notions of covalency, polarity, and multicentre delocalization into spatially resolved, quantitative descriptors, bridging classical chemical intuition with modern quantum theory and enabling a unified understanding of bonding across molecular systems.[19,20]

In parallel with theoretical developments, the past decade has witnessed a transformative shift toward data-driven molecular modelling, driven by advances in machine learning (ML) and the emergence of large, curated quantum chemical datasets.[21–23] Collections such as QM7,[24,25] QM8,[26,27] and particularly QM9[28] have provided standardized benchmarks

[a] *Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute of Science and Technology, 08028 Barcelona, Spain. E-mail: michal.michalski@irbbarcelona.org*
[b] *Faculty of Chemistry, University of Wroclaw, 50383 Wroclaw, Poland. E-mail: slawomir.berski@uwr.edu.pl*

for predicting molecular properties directly from atomic configurations. The QM9 dataset, which consists of numerous small organic molecules made up of elements up to the second row, provides high-quality DFT-calculated geometries. These datasets have enabled the training of ML models, ranging from kernel-based regressors[29,30] to message-passing neural networks,[31,32] that reproduce molecular properties such as total energies,[33] dipole moments,[34] and electronic descriptors[35] directly from atomic configurations at near-chemical accuracy but several orders of magnitude faster. More recently, the development of large-scale molecular repositories, such as ChEMBL,[36] PubChem,[37] ZINC database,[38] OMol25,[39] NAPROC,[40] and COCONUT[41,42] extending into the millions of chemical structures, has opened new opportunities for exploring the chemical space of up to medium-sized molecules.

The present work integrates QCT with ML to establish a data-driven framework for understanding chemical bonding at a large scale. Specifically, ELF-derived topological descriptors were computed for the QM9 dataset, allowing the extraction of bond-level information such as ELF basin populations ($\bar{N}$), interatomic distances (r), bond types, and chemical environments. Throughout this work, ELF basin populations corresponding to bonding basins are termed bond populations, using a bond-centric terminology to improve accessibility for audiences outside the QCT community. These descriptors were used to train and validate ML models capable of predicting bond populations directly from molecular geometry. By combining real-space bonding analysis with ML, this approach provides an interpretable pathway for mapping the chemical bonding across millions of molecules, bridging the gap between QCT and modern data science.

2 Materials and methods

The QM9 dataset[28] comprises 133 885 small organic molecules built from combinations of five chemical elements: carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and fluorine (F). Each entry is accompanied by quantum-chemically derived geometric, energetic, electronic, and thermodynamic properties, providing a rich library for data-driven exploration of molecular structure–property relationships. The dataset encompasses 6095 constitutional isomers, representing a chemically diverse yet systematically defined subset of the vast organic chemical space. Owing to its computational accuracy, internal consistency, and broad chemical coverage, QM9 has become one of the most widely used benchmark datasets for the development, validation, and benchmarking of ML models in theoretical chemistry.

To extend the QM9 dataset with topological information derived from real-space electron density, wavefunction (WFN) files were generated using the Gaussian 16 (G16, version C.01) program package.[43] Calculations were performed at the density functional theory (DFT) level employing the B3LYP functional[44–47] and the 6-31G(2df,p)[48–50] basis set, fully consistent with the computational protocol of the original QM9 study. The optimized molecular geometries available in QM9 were used directly as input to ensure comparability of electronic properties across the dataset. The resulting WFN files served as input for topological analysis of the ELF using the TopMod09 program package.[51] The ELF calculations were carried out on a cubic grid with a step size of 0.05 bohr.

The basin populations ($\bar{N}$) obtained from the integration of electron density for ELF-localization basins were processed to generate a ML dataset. The TopMod09 output files (.top) and corresponding wavefunction files (.wfn) were parsed to extract basin information and populations and to compute geometric and electronic descriptors for each ELF basin, corresponding to the description of chemical bonds. For each bonding attractor, the two nearest atomic centres were identified, and the corresponding interatomic distances were calculated. Valence bonding basins associated with the same atomic pair were merged, and the resulting data, comprising bond identity, defined as the pair of atoms forming a given chemical bond, electron (basin) population, bond length, and molecular identifier, were compiled into a unified dataset. Additionally, for each bond, the local chemical environment of the bonded atoms was described by identifying neighbouring atomic centres within a fixed spatial cut-off of 2.0 Å and enumerating their pairwise connections. The resulting environment labels capture the immediate bonding context around each atom, reflecting both coordination and compositional patterns within the molecule.

ML training was performed to predict ELF-derived $\bar{N}$ from geometric and bonding descriptors. The dataset was preprocessed by one-hot encoding the bond type and using the bond length as a continuous feature. Model hyperparameters were optimized through five-fold stratified cross-validation, in which stratification by bond type ensured balanced representation of different chemical bonds across training (80%) and test (20%) subsets, thereby preventing bias toward more frequent bond classes. Three regression algorithms were employed: (i) ridge regression, evaluated using regularization parameters α (0.1 to 10), with and without intercept fitting and with optional positivity constraints; (ii) gradient boosting random forest (GBRF) regression, implemented using the LightGBM framework with GPU acceleration, where the number of leaves (11 to 51), learning rate (0.001 to 0.1), and maximum tree depth (−1 to 10) were optimized; and (iii) a feed-forward neural network (FF-NN), constructed in Keras with the PyTorch backend, consisting of two hidden layers with neuron counts (512 to 5120) and dropout rates (0.25 to 0.75). The FF-NN was trained using the Adam optimizer with mean squared error (MSE) as the loss function and early stopping (patience 20 epochs) to prevent overfitting. After cross-validation, the production model was trained with the best-performing hyperparameter configurations in the full dataset. Model performance was assessed using mean absolute error (MAE). All calculations were carried out in Python 3.10 using scikit-learn 1.7.2,[52] LightGBM 4.6.0,[53] and Keras 3.11.3[54] with the PyTorch 3.6[55] backend.

3 Results and discussion

3.1 Exploratory data analysis for the QM9 dataset

The statistical analysis of bond-length distributions derived from the QM9 dataset reflects the diversity of covalent bonding

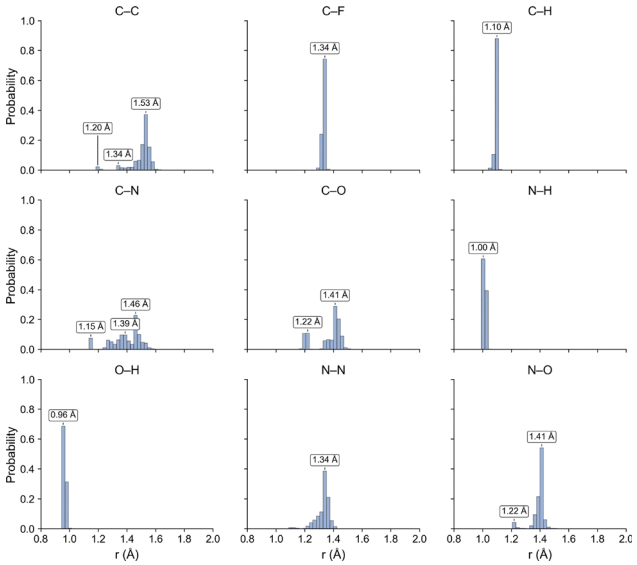

Fig. 1 Probability density distributions of equilibrium bond lengths (r) in the QM9 dataset computed at the B3LYP/6-31G(2df,p) level.

interactions (Fig. 1). Each histogram represents the normalized probability density function of equilibrium interatomic distances calculated at the B3LYP/6-31G(2df,p) level of theory, thereby capturing not only the most probable bond length but also the variability associated with different chemical environments. The C–H bond distribution exhibits a sharp, unimodal peak centred at 1.10 Å, with a narrow full width at half maximum, indicating a high degree of geometric uniformity across diverse molecular contexts. This value agrees closely with experimental measurements for hydrocarbons[56] and represents the theoretical equilibrium distance with a negligible systematic deviation. A similar distribution is observed for the O–H bond, with a probability maximum at 0.96 Å, reflecting the strong localization and limited structural flexibility of hydroxyl groups.

The C–C bond-length probability density displays a broader, asymmetric profile with a dominant maximum near 1.53 Å, consistent with the canonical single-bond distance observed in alkanes, but extending toward shorter values, where the tail of the distribution captures partial π-bonding and conjugated motifs.[57] The C–N and C–O bonds exhibit similar behaviour, with probability maxima at 1.46 Å and 1.41 Å, respectively, accompanied by secondary shoulders at shorter distances corresponding to amide, imine, and carbonyl substructures.[58] In both cases, the continuous nature of the distributions suggests a modulation of the bond length as a function of local hybridization and electronic delocalization rather than discrete bond classes. Additionally, both the C–N and C–O distributions display noticeable broadening, which can be attributed to their occurrence not only in typical functional groups but also within heterocyclic ring systems,[59] where geometric constraints and aromatic delocalization further modulate bond lengths. The C–F bonds form a narrow distribution centred at 1.34 Å, reflecting the rigidity of this strongly polarized covalent bond.[60] In contrast, the N–O bonds exhibit a distinct bimodal profile, with maxima near 1.22 Å and 1.41 Å, corresponding to N–O double bonds, as found in nitroso and nitro groups,[61] and N–O single bonds, as in nitrate and nitrite groups,[62] respectively. Although the N–N bonds in the QM9 dataset display an apparently symmetric, unimodal bond-length distribution centred at approximately 1.34 Å, from a chemical perspective, one would formally expect a bimodal or even trimodal distribution reflecting the coexistence of single, double, and triple N–N bonds. Typical equilibrium bond lengths for these interactions are well separated: N–N single bonds, as found in hydrazine-like motifs, occur near 1.45 Å,[63] N=N double bonds, characteristic of azo and diazene fragments, are observed around 1.25 Å,[64] and N≡N triple bonds, exemplified by dinitrogen, exhibit a much shorter distance of approximately 1.10 Å.[65] The absence of a clearly resolved trimodal structure in the QM9 statistics can be attributed to the limited chemical diversity of nitrogen-rich species in the dataset, as well as to the dominance of substituted or conjugated environments in which formal bond orders are partially delocalized. In such cases, resonance and hyperconjugation smear discrete bond-order classes into a continuous distribution of intermediate bond lengths. Consequently, the observed unimodal profile centred near 1.34 Å reflects an averaging over multiple bonding regimes rather than the exclusive presence of a single bond type.

For each bond type, the shape and dispersion of the probability density serve as statistical measures of structural variability and chemical-context dependence within the QM9 molecular space. Narrow distributions correspond to strongly localized σ-bonds with a minimal dependence on the surrounding molecular framework (C–H, N–H, O–H, and C–F), whereas broader or multimodal distributions reflect flexible bonding environments, resonance effects, and hybridization diversity (C–C, C–O, C–N, N–O, and N–N). The close correspondence between the modal bond lengths from QM9 and the experiment across all bond types confirms the geometric accuracy calculated at the B3LYP/6-31G(2df,p) level.

The analysis of $\bar{N}$ provides complementary insight into the electronic structure of chemical bonds within the QM9 dataset (Fig. 2). Each histogram represents the probability density function of the integrated electron population associated with a given bonding basin, as obtained from the topological ELF partitioning. The resulting distributions exhibit distinct and chemically interpretable features for each bond type, reflecting differences in covalent character, polarity, and electronic delocalization. The C–H, O–H, and N–H bonds display narrow, unimodal distributions with probability maxima near 2.0 e^-, consistent with two electrons localized in a strongly covalent σ-bond. The limited width of these peaks indicates a high degree of uniformity in electron sharing across the dataset, in agreement with the largely single-bonded character of hydrogen-containing groups.

The C–C bonds show a broader and slightly asymmetric probability distribution centred near 2.0 e^-, with a tail extending toward higher values up to 5.5 e^-, corresponding to π-delocalized systems and double or triple bond environments. This variation reflects the diversity of carbon–carbon bonding motifs in QM9, ranging from single σ-bonds in alkanes to conjugated and

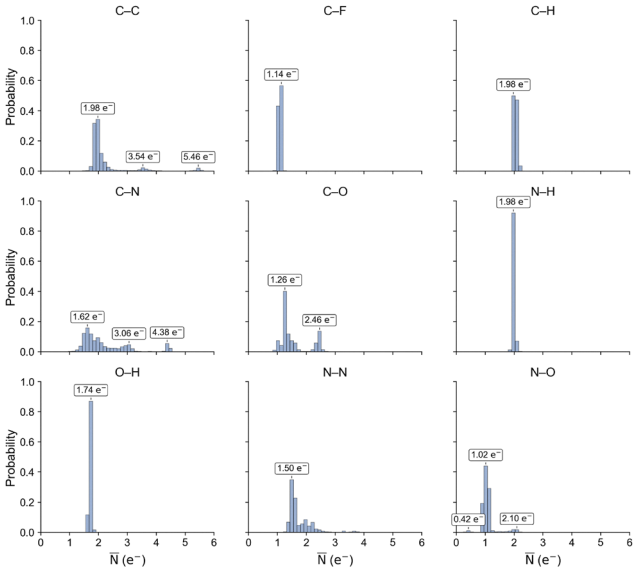

Fig. 2 Probability density distributions of ELF populations ($\bar{N}$) in the QM9 dataset computed at the B3LYP/6-31G(2df,p) level.

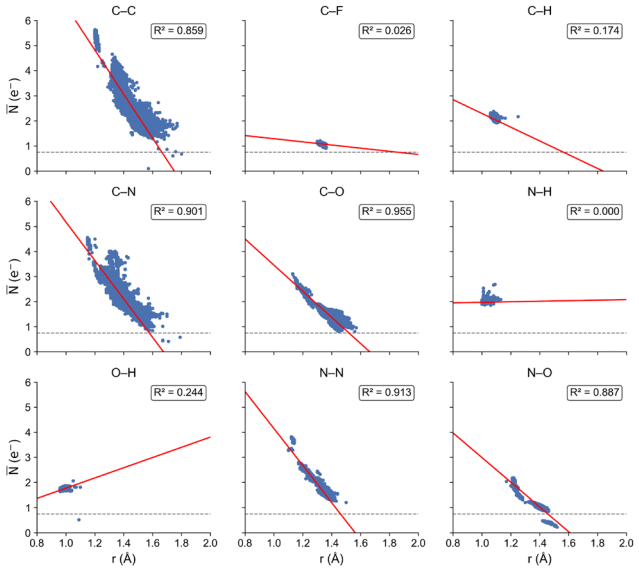

Fig. 3 Correlation between ELF population ($\bar{N}$) and bond length (r) in the QM9 dataset computed at the B3LYP/6-31G(2df,p) level. The dashed horizontal line indicates $\bar{N}$ = 0.75 e$^-$.

cumulene-like structures exhibiting increased electron localization between carbon centres. Similarly, the C–N and C–O bonds display broad, multimodal distributions spanning from 1.0 to 5.0 e$^-$, consistent with the coexistence of single and double bonds in both C–N and C–O and triple bonds in C–N. The presence of a significant population density below 2.0 e$^-$ indicates partial electron transfer toward the more electronegative atom, particularly pronounced for C–O bonds.

The C–F bonds yield a narrow but systematically shifted peak at 1.14 e$^-$, highlighting their strongly polarized nature and reduced shared-electron character due to the high electronegativity of fluorine. A similar trend is observed for N–O bonds, which exhibit maxima around 1.02 e$^-$, consistent with the mixed covalent-ionic character typical of nitro and oxo functional groups. Previously, one of us investigated the nature of the N–O bond, using the topological analysis of the ELF.[66] Such a bond, formally being a single bond, may exhibit a different topology of the ELF, which is characterized by a single disynaptic V(N,O) attractor, two monosynaptic V(N) and V(O) attractors, a single monosynaptic attractor V(N) and a lack of any valence attractors in the bonding region. The N–N bonds show a broader distribution extending from 1.0 to 3.0 e$^-$, reflecting their variable bond order and tendency toward delocalization. Across all bond types, the shape and dispersion of ELF population distributions provide statistical measures of electron-pair localization and its modulation by bond polarity and hybridization. High-probability peaks near 2.0 e$^-$ correspond to well-localized σ-bonds with electron sharing, whereas broader or multimodal distributions indicate variable bonding multiplicity, conjugation, or polarization effects.

The combined analysis of $\bar{N}$ and its correlation with bond lengths (r) provides a quantitative description of the relationship between geometric and electronic structures across the QM9 dataset. The probability distributions of $\bar{N}$ reveal the expected localization of approximately two electrons for most covalent σ-bonds, while the $\bar{N}(r)$ correlation plots demonstrate how the electron population varies with internuclear distance (Fig. 3). For each bond type, a linear regression was fitted to the $\bar{N}(r)$ data, and the corresponding coefficient of determination (R^2) was used as a measure of correlation strength. High R^2 values indicate a strong and chemically meaningful dependence of ELF population on bond length, whereas low or vanishing R^2 values identify bond types where electron localization is dominated by polarity or nonbonding contributions rather than by distance.

The strongest correlations were observed for C–C, C–N, C–O, N–N, and N–O bonds, all with $R^2 > 0.85$, reflecting the expected inverse relationship between basin population and internuclear distance: shorter bonds correspond to higher electron populations within the bonding basin. In particular, C–O (R^2 = 0.955) and N–N (R^2 = 0.913) display nearly linear trends, consistent with smooth transitions between single, double, and triple bonding regimes in these systems. The C–C and C–N correlations (R^2 = 0.86–0.90) similarly capture bond-order modulation arising from conjugation and hybridization effects. In contrast, bonds involving hydrogen: C–H, O–H, and N–H, exhibit narrow distributions of both r and $\bar{N}$, resulting in weak or negligible correlations ($R^2 < 0.25$). This behaviour reflects the nearly constant bond lengths and uniformly localized σ-bonds in hydrogen-containing fragments, where population fluctuations are minimal. The C–F bonds also display poor correlation (R^2 = 0.026) due to their strong polarity, as most of the electron density is concentrated near the fluorine atom, effectively decoupling the ELF population from geometric variation.

The occurrence of low ELF basin populations ($\bar{N} < 0.75$ e$^-$) can be attributed to electronic delocalization or topological artifacts inherent in the ELF partitioning procedure (Fig. 4). In many instances, the electron density that would normally be

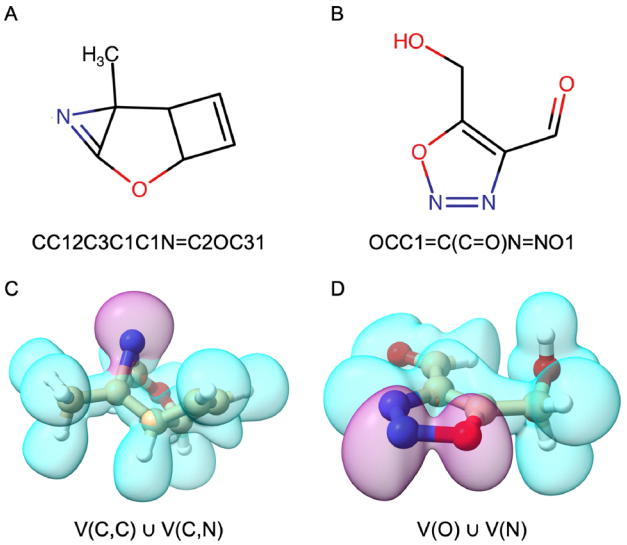

Fig. 4 Two-dimensional molecular structures (A, B) and corresponding ELF (isovalue = 0.65) domains (C, D) for QM9 molecules exhibiting low ELF basin populations (highlighted in purple). The examples shown correspond to IDs 066597 (A) and (C) and 131625 (B) and (D).

localized between two atoms is instead partially delocalized over two or more attractors, leading to the appearance of several weakly localized basins rather than a single well-defined bonding domain. Consequently, integration of the electron density over each individual basin yields a population smaller than the canonical value, even though the cumulative density within the entire delocalized region remains consistent with a shared-electron interaction. In strongly polarized bonds, such as C–N and N–O, the bonding electron density is displaced toward the more electronegative atom and partially incorporated into its adjacent lone-pair basin, resulting in a residual interatomic basin with a low population. A representative case in the QM9 dataset is entry 066597 (Fig. 4A and C), where the electron density is delocalized around the nitrogen atom across two C–N bonds and one C–C bond. The C–C bond is described by two disynaptic basins, V(C,C), with populations of 2.30 and 0.41 e^-, respectively, while the two C–N bonds correspond to V(C,N) basins with populations of 3.98 and 2.99 e^-. A similar effect is observed for QM9 entry 131625 (Fig. 4B and D), where multiple monosynaptic basins, V(N) and V(O), with populations from 0.23 e^- to 4.56 e^-, collectively describe the lone pairs on nitrogen and oxygen atoms, as well as the associated N–O bond. Thus, comparable effects are observed in geometrically constrained or highly strained systems, as well as in multicentre arrangements, where the ELF field exhibits diffuse or shallow maxima, reflecting numerical or topological segmentation rather than distinct localized electron pairs. Collectively, these observations indicate that basins with low populations arise in situations where electron localization is delocalized or not adequately captured by the synaptic assignment.

This correlation analysis serves as a filtering criterion for identifying chemically meaningful bonding basins for subsequent ML modelling. Bonds exhibiting high R^2 values correspond to shared-electron interactions that show well-defined relationships between bond length and electron population, whereas those with low R^2 values represent either highly polarized, ionic, or topologically ambiguous cases in which ELF attractors do not correspond to conventional two-centre bonds. The continuous and monotonic trends observed for C–C, C–N, and C–O systems further demonstrate that ELF populations encode a measure of bond order consistent with classical chemical intuition. Taken together, the geometric, electronic, and correlative analyses establish a statistical framework, linking molecular geometry, electron localization, and bond order across the QM9 dataset. These results provide the conditions for constructing data-driven models of bonding suitable for ML applications.

3.2 Regression models

The exploratory analysis revealed that ELF populations follow systematic trends with bond length, bond type, and local chemical environment, suggesting that these features should be used for the ML training process. Building on these observations, the next stage involved training regression models to predict ELF basin populations of C–C, C–N, C–O, N–N, and N–O bonds directly from molecular geometry. The primary objective was to evaluate the predictive capacity of different ML algorithms to reproduce the ELF-derived electron population from simple structural descriptors such as the bond type and bond length, and subsequently to assess how inclusion of local environmental features influences model accuracy (Fig. 5a and b).

Three representative regression models were selected to validate different levels of model complexity, interpretability, and performance. Ridge regression was included as a baseline to quantify the extent to which ELF populations can be described through simple linear relationships between geometric descriptors and electron localization. The GBRF model, based on an ensemble of decision trees, was chosen to capture nonlinear and nonadditive dependencies between features while retaining interpretability. Finally, the FF-NN model was employed as a flexible, high-capacity nonlinear model capable of learning complex, continuous mappings between molecular geometry and ELF populations. Together, these three methods provide a balanced comparison across linear, ensemble-based, and neural network paradigms, enabling systematic assessment of both the complexity of $\bar{N}$ prediction and the degree of nonlinearity required for accurate modelling. All models were trained and validated on the QM9 dataset using five-fold cross-validation, and performance was evaluated using the MAE metric.

In the first stage, models were trained using only the bond descriptors, bond identity and bond length, without incorporating any explicit information about the bond environment. Hyperparameter optimization was carried out *via* grid search. For ridge regression, the optimal configuration corresponded to a regularization strength of α = 10, with intercept fitting disabled and no positivity constraint. For the GBRF regression model, the lowest MAE was obtained for num_leaves = 51, learning_rate = 0.1, and max_depth = −1. For the FF-NN model,

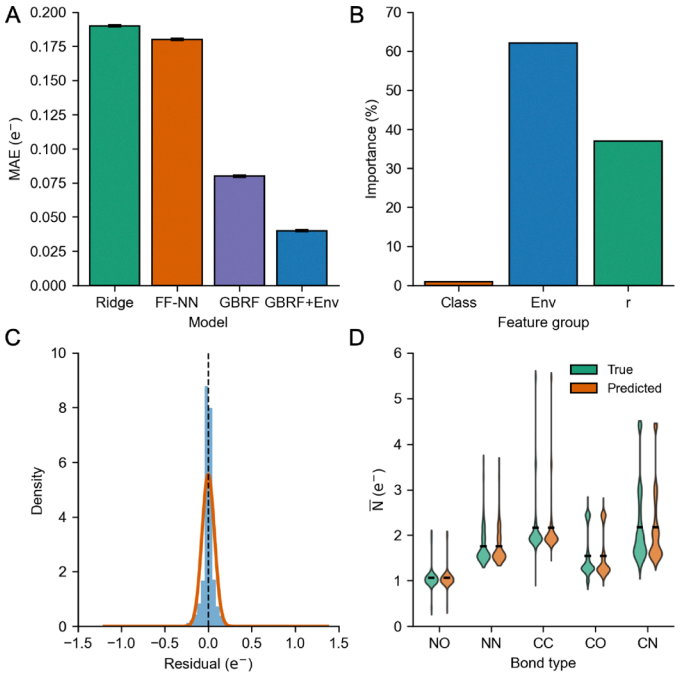

Fig. 5 Performance evaluation and residual analysis of trained ML models. (a) Mean absolute error (MAE) in the test set for four different model variants: ridge regression (ridge), feed-forward neural network (FF-NN), gradient boosting random forest regression (GBRF), and gradient boosting random forest with explicit environmental features (GBRF+Env). (b) Relative importance of the three input feature groups, grouped into Class (chemical identity), Env (environment descriptors), and *r* (bond length). (c) Histogram of test set residuals with a Gaussian fit (orange line). (d) Comparison of true and predicted ELF population distributions across bond types.

the optimal setup used 5120 neurons per hidden layer, a dropout rate of 0.25, and a learning rate of 0.001. Among these models (Fig. 5a), the ridge regression provided a baseline performance with an MAE of 0.19 ± 0.001 e^-, reflecting its limited ability to capture nonlinear relationships between geometric and electronic features. The FF-NN achieved an MAE of 0.18 ± 0.001 e^-; however, the difference relative to ridge regression does not support a meaningful improvement in predictive accuracy. The best performance was obtained with the GBRF model, which reached an MAE of 0.08 ± 0.001 e^-, demonstrating its capacity to model complex, non-additive dependencies between geometric descriptors and ELF populations.

In the second stage, the same regression framework was extended by including local chemical environment descriptors for each bond, representing the connectivity and elemental composition of neighbouring atoms. The GBRF model was optimized independently for this extended feature set, and the resulting best-performing parameters are num_leaves = 51, learning_rate = 0.1, and max_depth = −1. The incorporation of environmental information led to a marked improvement in predictive accuracy (Fig. 5a): whereas the GBRF model without environmental descriptors yielded an MAE of 0.08 ± 0.001 e^-, inclusion of environmental features reduced the error to 0.04 ± 0.001 e^-. This improvement highlights the role of the local bonding context in determining electron localization, as neighbouring atoms influence ELF-derived bond populations. Based on these findings, the GBRF regression model with environment descriptors was selected as the production model for further analysis.

To further evaluate the data efficiency and scalability of the trained models, learning curves were analysed as a function of training set size (Fig. S1 and S2, SI). These results explain how predictive performance evolves with increasing data availability. The GBRF model exhibits high data efficiency, achieving near-optimal performance even at small training fractions, with only marginal improvements upon increasing dataset size. This trend indicates that the model is capable of extracting the dominant structure–property relationships from relatively limited data. In contrast, the FF-NN model shows an improvement in accuracy by increasing the training set size up to approximately 50% of the dataset, beyond which performance saturates, suggesting that model capacity becomes the limiting factor rather than data availability. The ridge regression model displays limited scalability, with increasing error as the training set size grows, reflecting its inability to capture nonlinear dependencies in increasingly diverse chemical environments. Notably, the inclusion of environmental descriptors in the GBRF model leads to a consistent improvement in performance with increasing dataset size, highlighting the importance of local chemical context and its dependence on sufficient sampling of chemical space.

To gain deeper insight into the chemical interpretability of the trained model, detailed feature importance and residual analyses were conducted (Fig. 5b–d). The feature importance analysis revealed that the bond length constitutes the single most influential descriptor, contributing approximately 37% of the total predictive variance (Fig. 5b). The remaining 63% arises from descriptors representing the elemental identities of the

bonded atoms (1%) and the chemical environment (62%), which reflects modulation of electron density near the chemical bond. While the bond length is the most influential single descriptor, the local chemical environment plays a dominant role in determining electron localization within a bond. In effect, accurate prediction requires knowing not only how long the bond is but also where it exists within the molecular framework. Complementing these findings, the residual analysis (Fig. 5c) showed an approximately Gaussian error distribution centred near zero, with minimal systematic bias across bond classes, indicating no major systematic deficiencies, with remaining deviations likely arising from chemically complex bonding cases. To place these errors in the context, the distributions of target ELF populations were analysed (Fig. 5d). The true ELF populations span a broad range, from values near 0.5 e^- for highly polarized bonds to more than 5.5 e^- for the most electron-rich bonds. The predicted distributions closely follow the true distributions across all bond types, with good agreement in both medians and overall shapes, and only minor deviations are observed in the tails. In particular, the model reproduces the lower ELF populations observed in polarized bonds such as N–O and C–O, as well as the higher populations associated with more covalent bonds such as C–C and C–N, while also reflecting the substantial intra-type variability evident from the broad distributions. Taken together with MAE $\approx$ 0.04 e^-, these results confirm that the model performs well relative to the variability of the dataset and that the predictions remain reliable across the range of chemical bonds present in the QM9 dataset.

4 Conclusions

This study establishes a data-driven framework, integrating the QCT with ML to predict the electron population of chemical bonds. By combining ELF analysis with large-scale data of the QM9 dataset, we demonstrated that fundamental aspects of chemical bonding can be captured and predicted from simple geometric descriptors. The results may be summarised as follows:

(1) The exploratory data study of topological analysis of the ELF revealed consistency between bond lengths and populations across all bond types in the QM9 dataset. In addition, the $\bar{N}(r)$ correlation displays a uniform inverse trend, and shorter bonds carry higher basin populations, with strong linear correlations for C–C, C–N, C–O, N–N and N–O ($R^2 > 0.85$) and weak correlations for C–H, O–H, N–H, and C–F bonds, where limited geometric variability dominates.

(2) The regression analyses showed that ELF-derived bond populations can be predicted from the geometric structure alone, with nonlinear ML algorithms effectively capturing the dependencies between the bond length, bond type, and local environment. Among the tested models, GBRF achieved the best performance, reaching the lowest MAE when environmental features were included. Its balance between predictive accuracy and interpretability highlights GBRF as a transparent model for large-scale chemical space exploration.

(3) Beyond methodological considerations, the present work establishes the interplay between molecular geometry and the real-space bonding topology, which can be formalized through data-driven modelling. Although the current analysis is focused only on the QM9 dataset, the proposed framework is readily generalizable to other curated molecular databases, thereby enabling systematic investigations of chemical bonding across broader regions of chemical space. Also, the integration of geometrical descriptors within ML workflows facilitates the prediction and classification of chemical bonds across extensive molecular datasets, by reducing the simulation cost needed for computationally demanding conventional wavefunction-based analyses.

Author contributions

Conceptualisation, methodology, formal analysis, and writing – original draft preparation, M. M.; writing – review and editing, S. B.; visualisation, M. M.; and supervision, S. B. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6cp00029k.

All scripts used in this work are publicly available at https://github.com/Parecido/ELF-AI. The repository contains Python scripts for dataset preparation, cross-validation, hyperparameter optimization, and training of the production model. Example wavefunction (.wfn) and topology (.top) files are also provided to illustrate the data format and workflow. The raw data supporting the conclusions will be made available by the authors on request.

Acknowledgements

The authors are grateful to the PLGrid Infrastructure and the Wroclaw Centre for Networking and Supercomputing for allocation of computing time.

Notes and references

1 L. Zhao, W. H. E. Schwarz and G. Frenking, *Nat. Rev. Chem.*, 2019, **3**, 35–47.
2 S. Esposito and A. Naddeo, *Adv. Hist. Stud.*, 2014, **03**, 229–257.
3 W. Heitler and F. London, *Z. Phys.*, 1927, **44**, 455–472.

4 P. L. A. Popelier, in *On Quantum Chemical Topology*, ed. R. Chauvin, C. Lepetit, B. Silvi and E. Alikhani, Springer International Publishing, Cham, 2016, pp. 23–52.
5 P. L. A. Popelier, *Drug Design Strategies Computational Techniques and Applications*, The Royal Society of Chemistry, 2012.
6 R. F. W. Bader, *Atoms in Molecules: A Quantum Theory*, Oxford University Press, USA/Oxford, 1994.
7 R. F. W. Bader, J. Hernández-Trujillo and F. Cortés-Guzmán, *J. Comput. Chem.*, 2007, **28**, 4–14.
8 A. D. Becke and K. E. Edgecombe, *J. Chem. Phys.*, 1990, **92**, 5397–5403.
9 B. Silvi and A. Savin, *Nature*, 1994, **371**, 683–686.
10 B. Silvi and R. J. Gillespie, *The ELF Topological Analysis Contribution to Conceptual Chemistry and Phenomenological Models*, John Wiley & Sons, Ltd, 2007, ch. 6, pp. 141–162.
11 Y. Grin, A. Savin and B. Silvi, *The ELF Perspective of chemical bonding*, John Wiley & Sons, Ltd, 2014, ch. 10, pp. 345–382.
12 B. Silvi, *The Relevance of the ELF Topological Approach to the Lewis, Kossel, and Langmuir Bond Model*, 2015, vol. 170.
13 M. Kohout, *Int. J. Quantum Chem.*, 2004, **97**, 651–658.
14 B. Silvi, I. Fourré and M. E. Alikhani, *Monatsh. Chem.*, 2005, **136**, 855–879.
15 B. Silvi, *J. Mol. Struct.*, 2002, **614**, 3–10.
16 A. Savin, B. Silvi and F. Colonna, *100 Years of CSC in the Pages of CJC*, 2017, vol. 01, pp. 1088–1096.
17 X. Krokidis, S. Noury and B. Silvi, *J. Phys. Chem. A*, 1997, **101**, 7277–7282.
18 R. Thom, *Structural Stability and Morphogenesis: An Outline of a General Theory of Models*, W. A. Benjamin, New York, 1975.
19 J. Poater, M. Duran, M. Solà and B. Silvi, *Chem. Rev.*, 2005, **105**, 3911–3947.
20 B. Silvi, M. Alikhani, C. Lepetit and R. Chauvin, *Topological Approaches of the Bonding in Conceptual Chemistry*, 2016.
21 R. Xia and S. Kais, *Nat. Commun.*, 2018, **9**, 4195.
22 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chem. Rev.*, 2021, **121**, 10142–10186.
23 B. Huang and O. A. von Lilienfeld, *Chem. Rev.*, 2021, **121**, 10001–10036.
24 L. C. Blum and J.-L. Reymond, *J. Am. Chem. Soc.*, 2009, **131**, 8732.
25 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.
26 L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.
27 R. Ramakrishnan, M. Hartmann, E. Tapavicza and O. A. von Lilienfeld, *J. Chem. Phys.*, 2015, **143**, 084111.
28 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
29 M. J. Willatt, F. Musil and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2018, **20**, 29661–29668.
30 A. S. Christensen, F. A. Faber and O. A. von Lilienfeld, *J. Chem. Phys.*, 2019, **150**, 064105.
31 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, Proceedings of the 34th International Conference on Machine Learning, 2017, vol. 70, pp. 1263–1272.
32 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, *Nat. Commun.*, 2022, **13**, 2453.
33 A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth and B. Kozinsky, *Nat. Commun.*, 2023, **14**, 579.
34 O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.*, 2019, **15**, 3678–3693.
35 A. M. Lewis, A. Grisafi, M. Ceriotti and M. Rossi, *J. Chem. Theory Comput.*, 2021, **17**, 7203–7214.
36 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.
37 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. Shoemaker, P. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. Bolton, *Nucleic Acids Res.*, 2025, **53**, D1516–D1525.
38 B. I. Tingle, K. G. Tang, M. Castanon, J. J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun, Y. S. Moroz and J. J. Irwin, *J. Chem. Inf. Model.*, 2023, **63**, 1166–1176.
39 D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman, N. C. Frey, X. Fu, V. Gharakhanyan, A. S. Krishnapriyan, J. A. Rackers, S. Raja, A. Rizvi, A. S. Rosen, Z. Ulissi, S. Vargas, C. L. Zitnick, S. M. Blau and B. M. Wood, *The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models, arXiv*, 2025, preprint, arxiv:2505.08762, DOI: 10.48550/arxiv.2505.08762 https://arxiv.org/abs/2505.08762.
40 J. F. Avellaneda-Tamayo, N. A. Agudo-Muñoz, J. E. Sánchez-Galán, J. L. López-Pérez and J. L. Medina-Franco, *J. Nat. Prod.*, 2024, **87**, 2216–2229.
41 M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik and C. Steinbeck, *J. Cheminf.*, 2021, **13**, 2.
42 V. Chandrasekhar, K. Rajan, S. R. S. Kanakam, N. Sharma, V. Weißenborn, J. Schaub and C. Steinbeck, *Nucleic Acids Res.*, 2025, **53**, D634–D643.
43 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 A.03*, Gaussian Inc., Wallingford CT, 2016.
44 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.

45 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
46 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
47 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
48 W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257–2261.
49 M. J. Frisch, J. A. Pople and J. S. Binkley, *J. Chem. Phys.*, 1984, **80**, 3265–3269.
50 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, *J. Chem. Phys.*, 1980, **72**, 650–654.
51 S. Noury, X. Krokidis, F. Fuster and B. Silvi, *Comput. Chem.*, 1999, **23**, 597–604.
52 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
53 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, pp. 3149–3157.
54 F. Chollet *et al.*, *Keras*, https://keras.io, 2015.
55 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, *PyTorch: an imperative style, high-performance deep learning library*, Curran Associates Inc., Red Hook, NY, USA, 2019.
56 F. H. Allen and I. J. Bruno, *Acta Crystallogr., Sect. B:Struct. Sci.*, 2010, **66**, 380–386.
57 T. Y. Nikolaienko, V. S. Chuiko and L. A. Bulavin, *Comput. Theor. Chem.*, 2019, **1163**, 112508.
58 L. J. C. van der Zee, J. Hofman, S. Mathew, A. de Visser, E. Brück, B. de Bruin and J. Chris Slootweg, *Chem. – Eur. J.*, 2025, **31**, e202403885.
59 F. H. Allen, *Acta Crystallogr., Sect. B:Struct. Sci.*, 2002, **58**, 380–388.
60 T. Ribbeck, C. Kerpen, D. Löw, A. E. Sedykh, K. Müller-Buschbaum, N. V. Ignat'ev and M. Finze, *J. Fluorine Chem.*, 2019, **219**, 70–78.
61 J. Nicolas, Y. Guillaneuf, D. Bertin, D. Gigmes and B. Charleux, *Polymer Science: A Comprehensive Reference*, Elsevier, Amsterdam, 2012, pp. 277–350.
62 H. Oberhammer, *J. Mol. Struct.*, 2002, **605**, 177–185.
63 R. Coufal and J. Vohlídal, *Sci. Rep.*, 2023, **13**, 17383.
64 E. M. Novikov, J. Guillen Campos, J. Read de Alaniz, M. S. Fonari and T. V. Timofeeva, *Acta Crystallogr., Sect. E:Crystallogr. Commun.*, 2024, **80**, 867–872.
65 D. Laniel, F. Trybel, A. Aslandukov, J. Spender, U. Ranieri, T. Fedotenko, K. Glazyrin, E. L. Bright, S. Chariton, V. B. Prakapenka, I. A. Abrikosov, L. Dubrovinsky and N. Dubrovinskaia, *Nat. Commun.*, 2023, **14**, 6207.
66 S. Berski and A. J. Gordon, in *Diversity of the Nature of the Nitrogen-Oxygen Bond in Inorganic and Organic Nitrites in the Light of Topological Analysis of Electron Localisation Function (ELF)*, ed. R. Chauvin, C. Lepetit, B. Silvi and E. Alikhani, Springer International Publishing, Cham, 2016, pp. 529–551.