doi:10.14311/ppt.2025.2.152
Plasma Physics and Technology **12**(2):152–157, 2025

RAPID PREDICTION OF INSULATING GAS PERFORMANCE: LEVERAGING XTB CALCULATIONS FOR HIGH-THROUGHPUT SCREENING OF SF_6 ALTERNATIVES

P. Liu, W. Cao, Y. Yao, B. Zhang, X. Li*

State Key Laboratory of Electrical Insulation and Power Equipment, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi Province, People's Republic of China

* xwli@mail.xjtu.edu.cn

Abstract. Addressing the environmental challenge posed by sulfur hexafluoride (SF_6) due to its extremely high Global Warming Potential (GWP), the development of eco-friendly alternative gases is imperative. Efficient prediction of key gas properties is crucial for virtual screening and generative model optimization, but traditional quantum chemical calculations for molecular descriptors are time-consuming for high-throughput screening. This study proposes and validates a rapid modeling framework using xTB semi-empirical quantum chemical calculations to efficiently compute physicochemical parameters, which provide the basis for core gas properties such as dielectric strength, boiling point, and GWP. Testing confirms the xTB method offers both high speed and acceptable accuracy, which establishes a foundation for virtual screening, accelerating the discovery of environmentally benign insulating gases.

Keywords: Insulating Gas, xTB Semi-Empirical Method, High-throughput screening, Dielectric strength.

1. Introduction

Sulfur hexafluoride (SF_6), with its excellent insulation and arc-quenching properties, has been widely used in electrical equipment for over half a century [1, 2]. However, SF_6 is an extremely stable man-made greenhouse gas with a Global Warming Potential (GWP) approximately 25,200 times that of carbon dioxide and an atmospheric lifetime of 3,200 years [3, 4]. With increasing global concern about climate change, finding and developing low-GWP, high-performance alternatives to SF_6 has become an urgent task for the power industry and academia.

The methods for discovering new insulating gas molecules have undergone continuous evolution. Early stages relied heavily on expert chemical intuition and extensive trial-and-error experimentation. For instance, researchers at ABB and GE synthesized a series of promising fluorine-containing compounds and tested their breakdown voltage, boiling point, and stability one by one, ultimately discovering promising alternative gases such as C5-fluoroketone ($C_5F_{10}O$) [5] and C4-fluoronitrile (C_4F_7N) [6].

However, this approach is not only costly and time-consuming, but its exploratory scope is also severely limited by known chemical reaction pathways and synthesis experience. With the advancement of computational science, virtual screening technology has emerged. It predicts molecular properties using computational chemistry methods to screen potential candidates from large-scale molecular libraries. For example, X. Li *et al.* performed an initial screening of the billion-scale small molecule library from ZINC based on boiling point constraints. They then conducted high-throughput DFT calculations on the resulting 1,248 candidate molecules to predict boiling point, dielectric strength, and GWP, ultimately identifying 41 compounds as potential SF_6 substitutes [7].

A key bottleneck for both virtual screening lies in the high-throughput, high-efficiency prediction of the critical properties of candidate molecules. Dielectric strength, the core metric for a gas's insulation capability, is crucial to predict accurately. Previous research has largely relied on quantum chemistry software based on Density Functional Theory (DFT), such as Gaussian, to obtain molecular electronic structure information (e.g., molecular polarizability, HOMO-LUMO gap) as descriptors for building performance prediction models. H. Sun *et al.* used the B3LYP method to calculate seven molecular descriptors, including polarizability, electronegativity, and positive surface area, to construct a dielectric strength prediction model [8]. L. Luo *et al.* used the M06-2X method to calculate five molecular descriptors, including the Lowest Unoccupied Molecular Orbital (LUMO) energy, molecular volume, and polarizability, to build their prediction model [9]. Although the descriptors used in the two papers differ, both employed GAUSSIAN for DFT calculations. While these first-principles DFT methods offer high accuracy, their computational cost is extremely high, with calculations for a single molecule often taking several minutes or even longer. This limits the speed of screening, making it difficult to apply to large-scale screening tasks involving millions or even more molecules.

In fact, semi-empirical methods, represented by

xTB [10], have already shown great potential in high-throughput screening across various electrical and energy materials fields due to their excellent computational efficiency. In the field of organic semiconductors, L. Wilbraham *et al.* and M. O. Faruque *et al.* used the xTB series of methods to rapidly calculate the optoelectronic properties of conjugated polymers [11] and the charge carrier mobility of organic molecular semiconductors [12], respectively, greatly accelerating the discovery process for new materials. In perovskite solar cell research, S. Raaijmakers *et al.* achieved accurate dynamic simulations of the phase stability of $CsPbX_3$ perovskites by fine-tuning GFN1-xTB parameters [13]. In the energy storage sector, GFN2-xTB has been used in molecular dynamics simulations to investigate the structure of carbonate-based electrolytes and ion-solvent interactions in lithium/sodium batteries [14]. These studies fully demonstrate the effectiveness of xTB in handling complex molecular systems and predicting their electrical and optical properties. However, a literature review reveals that applying the xTB method to predict the dielectric strength of insulating gases is a significant, unexplored research gap.

Therefore, this paper aims to fill this gap by introducing the xTB semi-empirical quantum chemistry calculation method to the rapid prediction of insulating gas properties for the first time. We will construct a prediction framework based on descriptors calculated by xTB and conduct a comprehensive comparison with the traditional Gaussian calculation method in terms of accuracy and speed. This will validate the effectiveness and reliability of the proposed method for achieving high-throughput screening of environmentally friendly insulating gases.

2. Method

This study aims to build machine learning models to predict the dielectric strength of gases. To ensure a scientific and objective evaluation, experimentally measured dielectric strength values were used as the prediction target. The selection of input features (i.e., quantum chemical descriptors) was guided by established physical models, which indicate that parameters such as molecular polarizability, ionization potential, electron affinity, and the HOMO-LUMO gap are closely related to a gas's insulating capability [8, 9]. Therefore, our strategy was to first compute these physically meaningful descriptors using two contrasting quantum chemistry methods: high-precision DFT (GAUSSIAN) and the high-speed semi-empirical approach (xTB). Subsequently, an automated machine learning framework was utilized to construct predictive models from each set of descriptors to ascertain the comparative performance of the two methods in terms of accuracy and speed.

2.1. High-Precision Descriptor Calculation using Gaussian and Multiwfn

Gaussian [15] is a quantum chemistry software package widely used in both academic and industrial sectors. In this study, we employed first-principles Density Functional Theory (DFT) for our calculations. The theoretical foundation of DFT is the solution of the Kohn-Sham equations, which, proceeding from fundamental physical constants, can in principle accurately describe the properties of a system.

Specifically, we selected B3LYP, a classic hybrid functional. This functional has been demonstrated to be reliable and accurate in the study of various chemical systems by incorporating a proportion of exact Hartree-Fock exchange energy. For the basis set, we chose 6-31+g(d,p), a relatively complete basis set capable of providing a flexible description of molecular orbitals.

Our computational workflow consisted of two steps: first, we performed geometry optimization and frequency calculations on the molecules using Gaussian to ensure that the stable conformations obtained were true minima on the potential energy surface. This step also generated the corresponding wavefunction information files (.fchk). Subsequently, the wavefunction analysis software Multiwfn[16, 17] was used to post-process the Gaussian output files to calculate and extract a rich set of quantum chemical descriptors.

Through this computational chain, we systematically obtained a total of 11 scalar descriptors, which are organized into the following 3 categories:

1. **Energy and Structure (2 descriptors)**: The Self-Consistent Field (SCF) energy and the final gradient norm of the molecule.

2. **Energy Components (3 descriptors)**: Energy terms within the DFT framework, including electron kinetic energy, Coulomb energy, and exchange-correlation energy.

3. **Electronic Properties (6 descriptors)**:

   a. **Orbital Energies**: The energies of the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO), as well as the HOMO-LUMO gap.

   b. **Ionization Potential (IP) and Electron Affinity (EA)**: Calculated via the ΔSCF method ($IP = E_{\text{cation}} - E_{\text{neutral}}$; $EA = E_{\text{neutral}} - E_{\text{anion}}$).

   c. **Molecular Polarizability (α)**: The static polarizability of the molecule was calculated using the Polar keyword.

This combined approach of using Gaussian and Multiwfn not only yields high-precision molecular conformations and energies but also provides a series of quantum chemical descriptors that are physically meaningful and crucial for machine learning modeling through detailed wavefunction post-processing.

2.2. Rapid Descriptor Calculation using xTB

To overcome the efficiency limitations of DFT, we introduced the eXtended Tight-Binding (xTB) method [10] for rapid calculations. xTB is a suite of semi-empirical tight-binding quantum chemistry methods developed by the Grimme group. Its theoretical core can be viewed as a systematic simplification and approximation of the DFT total energy.

Unlike ab initio methods, semi-empirical methods achieve a computational speedup of 2–3 orders of magnitude (100–1000 times) compared to traditional DFT by neglecting or parameterizing the most time-consuming terms, such as the two-electron repulsion integrals, and introducing parameters fitted to high-precision reference data to compensate for the errors introduced by approximation, thereby trading some theoretical accuracy for significantly enhanced efficiency.

We specifically employed the GFN1-xTB method, as implemented in the xtb program package (version 6.4.1) from the Grimme group [18]. This method has been "purpose-driven" parameterized and is explicitly designed to reliably reproduce molecular geometries (G), vibrational frequencies (F), and noncovalent interactions (N). Its Hamiltonian treats intramolecular and intermolecular interactions through a simplified electrostatic model and an empirical dispersion correction. These design features enable it to rapidly obtain geometries that are in good agreement with high-precision methods.

Based on the stable conformations optimized with GFN1-xTB, we employed a systematic procedure to calculate molecular descriptors. In particular, for the key property of molecular polarizability, we used a computational chain consisting of the xtb4stda and stda programs to achieve a complete workflow from wavefunction generation to optical property calculation. This process first generates xTB wavefunction information using xtb4stda, followed by an excited-state calculation performed by the stda program to obtain the static polarizability by calculating at a very large wavelength. Through this procedure, a total of 11 scalar descriptors were obtained:

1. **Energy and Structure (2 descriptors)**: The Self-Consistent Charge (SCC) energy, and the final gradient norm of the molecule.
2. **Energy Components (3 descriptors)**: Energy terms within the semi-empirical framework, including electrostatic energy, repulsion energy, and dispersion energy.
3. **Electronic Properties (6 descriptors)**:
   a. **Orbital Energies**: The energies of the HOMO and LUMO, the HOMO-LUMO gap.
   b. **Ionization Potential (IP) and Electron Affinity (EA)**: Directly extracted from the xTB calculation results.
   c. **Molecular Polarizability (α)**: The static polarizability of the molecule was obtained through excited-state calculations using the xtb4stda and stda program chain.

This method not only leverages the high efficiency of xTB in structure optimization but also systematically investigates its performance in terms of electronic structure and energy composition through a dedicated computational chain and detailed output parsing, providing a rich and multi-dimensional feature set for subsequent machine learning modeling.

2.3. Automated Machine Learning (AutoGluon)

To ensure the objectivity of model evaluation and to eliminate biases introduced by manual hyperparameter tuning, this study employed AutoGluon [19], an open-source automated machine learning (AutoML) framework. The core advantage of AutoGluon lies in its high degree of automation, enabling it to automatically handle a series of complex processes including data preprocessing, model selection, and hyperparameter optimization. It does not rely on a single model but rather constructs high-performance and robust ensemble predictors by automatically integrating multiple base models, including gradient boosting trees (e.g., LightGBM, CatBoost) and deep neural networks, through advanced multi-layer stacking and bagging techniques.

In this study, we designed three sets of control experiments to systematically evaluate the impact of different descriptor sets on the predictive performance of the models. The specifics are as follows:

1. **Group 1 (Baseline)**: Used only molecular topological descriptors generated by the open-source cheminformatics library RDKit.
2. **Group 2**: Supplemented the RDKit descriptors with high-precision quantum chemical descriptors calculated by GAUSSIAN.
3. **Group 3**: Supplemented the RDKit descriptors with rapid quantum chemical descriptors calculated by xTB.

Using experimentally measured dielectric strength as the prediction target, we modeled each of the three feature sets using AutoGluon. To ensure a fair comparison, all modeling tasks were conducted under uniform conditions: the model quality was preset to high quality to construct high-performance ensemble models, and the training time for each task was capped at 1800 seconds. The dataset was partitioned into a training set (161 molecules) and a test set (40 molecules), which corresponds to an approximate 80:20 ratio. To strictly prevent data leakage, the training, validation, and optimization of the models were performed entirely within the training set using cross-validation. The test set was not involved in any training process and was used solely for final performance evaluation. Finally, by comparing the coefficient of determination (R^2), root mean square error (RMSE), mean squared

Molecule	Gau. (s)	xTB (s)	Rate
CH_3F	8.67	0.133	65.188
c-C_7F_{14}	403.376	1.229	328.215
Average	**44.459**	**0.338**	**131.536**

Table 1. Comparison of calculation times (s) between the Gaussian and xTB methods.

Model	R^2	RMSE	MAE
Rdkit	0.6810	0.30	0.24
Rdkit+xTB	0.7566	0.26	0.19
Rdkit+GAUSSIAN	0.7822	0.24	0.19

Table 2. Performance comparison of the prediction models.

error (MSE), and mean absolute error (MAE) of each model on the independent test set, we quantitatively assessed the impact of introducing different quantum chemical descriptors on prediction accuracy.

3. Results

This study aims to systematically evaluate the relative performance of the xTB semi-empirical method in terms of computational efficiency and prediction accuracy. To this end, a dataset of 201 molecules was established by compiling molecules with known dielectric strengths from two literature sources [20, 21], supplemented by our own experimental measurements. All dielectric strength values in the dataset correspond to standard conditions (atmospheric pressure at 1 atm and room temperature at 298 K) to ensure data consistency. The calculation time is shown in Table 1. Using the Gaussian B3LYP/6-31+g(d,p) method, the average time for a complete geometry optimization, frequency analysis, and property calculation for a single molecule was 44.459 seconds. In contrast, the xTB/GFN1 method required only 0.338 seconds, resulting in an average increase in computational efficiency of 131-fold.

The analysis of computational efficiency reveals a clear pattern: the computational speed of the xTB method is significantly superior to the traditional Gaussian method. Further analysis indicates that this efficiency advantage is positively correlated with molecular complexity. For the relatively simple molecule CH_3F, the speed-up factor is 65, whereas for the complex molecule c-C_7F_{14}, the factor increases to 328. This order-of-magnitude improvement in computational performance effectively overcomes the limitation of traditional quantum chemistry calculations being too time-consuming for large-scale screening, thus making high-throughput virtual screening technically feasible.

In terms of prediction accuracy, this study utilized three sets of descriptors and the AutoGluon automated machine learning framework to construct predictive models for dielectric strength. The performance comparison of the three prediction models is shown in Table 2. The results show that the model built solely on Rdkit topological descriptors achieved a coefficient of determination (R^2) of 0.6810 on the test set. Upon incorporating quantum chemical descriptors calculated by xTB, the model's R^2 increased

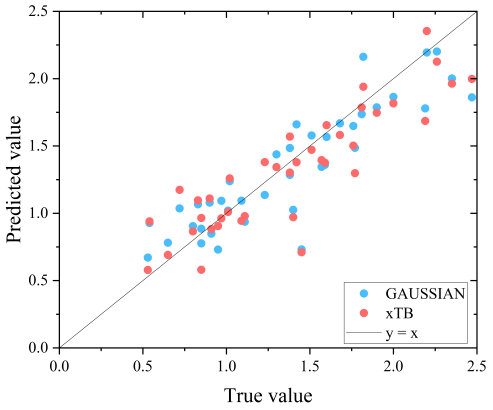

Figure 1. Parity plot for the dielectric strength predictions from the Gaussian and xTB models.

to 0.7566. For comparison, introducing descriptors calculated by Gaussian resulted in an R^2 of 0.7822.

Figure 1 shows the comparison of the predictions of dielectric strength by Gaussian and xTB models. The results indicate that although the accuracy of the predictive model based on xTB descriptors is slightly lower than that based on Gaussian descriptors, the difference in their R^2 values is small. Within the automated machine learning framework, the model achieved an R^2 of 0.75, which generally meets the requirements for predictive models in virtual screening and molecular generation tasks. It is noteworthy that this result was obtained under automated hyperparameter tuning without manual intervention; subsequent manual feature selection and model parameter fine-tuning are expected to further enhance its predictive accuracy. Considering the more than 100-fold increase in computational efficiency brought by the xTB method, the minor loss in prediction accuracy is well within an acceptable range. Therefore, the xTB method demonstrates significant application value in the early stages of screening, which require rapid exploration of vast chemical spaces.

4. Conclusions

This study has successfully constructed and validated a computational framework for the rapid prediction of the performance of environmentally friendly insulating gases. By substituting traditional, time-consuming Gaussian ab initio methods with xTB semi-empirical quantum chemistry calculations to obtain molecu-

lar descriptors and subsequently building a dielectric strength prediction model, this framework achieves a substantial improvement in computational efficiency within an acceptable range of accuracy loss for the prediction model.

The main features of this framework are as follows:

1. **Efficiency**: Compared to traditional Gaussian DFT calculations, the xTB method reduces the average descriptor calculation time for a single molecule by two orders of magnitude, providing technical feasibility for large-scale virtual screening and real-time molecular generation.

2. **Reliability**: The predictive model built on xTB descriptors, while slightly less accurate than the model based on Gaussian descriptors, still maintains a good level of performance ($R^2 > 0.75$), and if hyperparameters are further adjusted manually to improve prediction accuracy, it can be used for large-scale virtual screening.

3. **Physical Basis**: The key descriptors calculated by xTB, such as molecular polarizability and the HOMO-LUMO energy gap, are strongly correlated with the intrinsic physical mechanisms of gas insulation, confirming the physical reliability of the method.

In summary, the proposed framework of "rapid xTB calculation" offers an effective and viable solution to the high-throughput prediction bottleneck in the current research and development of environmentally friendly insulating gases. This framework not only accelerates the screening process of existing molecular libraries, but can also be embedded as a core engine in inverse molecular design workflows, thereby supporting the efficient discovery and optimization of new molecular structures and paving the way for finding ideal substitutes for SF_6.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (52422707).

References

[1] C. T. Dervos and P. Vassiliou. Sulfur hexafluoride (SF_6): global environmental effects and toxic byproduct formation. *Journal of the Air & Waste Management Association*, 50(1):137–141, 2000. doi:10.1080/10473289.2000.10463996.

[2] A. Xiao, J. G. Owens, J. Bonk, et al. Environmentally friendly insulating gases as SF_6 alternatives for power utilities. In *2019 2nd International Conference on Electrical Materials and Power Equipment (ICEMPE)*, pages 42–48. IEEE, 2019. doi:10.1109/ICEMPE.2019.8727308.

[3] V. Masson-Delmotte et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2(1):2391, 2021. arXiv:https://www.ipcc.ch/report/ar6/wg1/.

[4] J. G. Owens. Greenhouse gas emission reductions through use of a sustainable alternative to SF_6. In *2016 IEEE Electrical Insulation Conference (EIC)*, pages 535–538. IEEE, 2016. doi:10.1109/EIC.2016.7548658.

[5] M. Hyrenbach and S. Zache. Alternative insulation gas for medium-voltage switchgear. In *2016 Petroleum and Chemical Industry Conference Europe (PCIC Europe)*, pages 1–9. IEEE, 2016. doi:10.1109/PCICEurope.2016.7604648.

[6] Y. Kieffel, F. Biquez, D. Vigouroux, et al. Characteristics of g^3–an alternative to SF6. *CIRED 24*, 2017(1):54–57, 2017. doi:10.1109/ICD.2016.7547757.

[7] X. Li, D. Sun, Y. Zhou, et al. Virtual screening of new high voltage insulating gases as potential candidates for SF6 replacement. In *Frontier Academic Forum of Electrical Engineering*, pages 739–752. Springer, 2022. doi:10.1007/978-981-99-3408-9_64.

[8] H. Sun, L. Liang, C. Wang, et al. Prediction of the electrical strength and boiling temperature of the substitutes for greenhouse gas SF_6 using neural network and random forest. *IEEE Access*, 8:124204–124216, 2020. doi:10.1109/ACCESS.2020.3004519.

[9] L. Luo, S. Yang, Z. Yang, et al. A prediction model for electrical strength of gaseous medium based on molecular reactivity descriptors and machine learning method. *Journal of Molecular Modeling*, 31(2):53, 2025. doi:10.1007/s00894-024-06254-y.

[10] S. Grimme, C. Bannwarth, and P. Shushkov. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements ($z = 1 - -86$). *Journal of chemical theory and computation*, 13(5):1989–2009, 2017. doi:10.1021/acs.jctc.7b00118.

[11] L. Wilbraham, E. Berardo, L. Turcani, et al. High-throughput screening approach for the optoelectronic properties of conjugated polymers. *Journal of chemical information and modeling*, 58(12):2450–2459, 2018. doi:10.1021/acs.jcim.8b00256.

[12] M. O. Faruque, S. Akter, D. K. Limbu, et al. High-throughput screening, crystal structure prediction, and carrier mobility calculations of organic molecular semiconductors as hole transport layer materials in perovskite solar cells. *Crystal Growth & Design*, 24(21):8950–8960, 2024. doi:10.1021/acs.cgd.4c00965.

[13] S. Raaijmakers, M. Pols, J. M. Vicent-Luna, and S. Tao. Refined GFN1-xTB parameters for engineering phase-stable $CsPbX_3$ perovskites. *The Journal of Physical Chemistry C*, 126(22):9587–9596, 2022. doi:10.1021/acs.jpcc.2c02412.

[14] P. Wróbel and A. Eilmes. Effects of Me–solvent interactions on the structure and infrared spectra of MeTFSI (Me= Li, Na) solutions in carbonate solvents—a test of the GFN2-xTB approach in molecular dynamics simulations. *Molecules*, 28(18):6736, 2023. doi:10.3390/molecules28186736.

[15] M. J. Frisch, G. W. Trucks, H. B. Schlegel, et al. Gaussian~16 Revision B.01, 2016. https://gaussian.com/gaussian16/.

[16] T. Lu and F. Chen. Multiwfn: A multifunctional wavefunction analyzer. *Journal of computational chemistry*, 33(5):580–592, 2012. doi:10.1002/jcc.22885.

[17] T. Lu. A comprehensive electron wavefunction analysis toolbox for chemists, Multiwfn. *The Journal of chemical physics*, 161(8), 2024. doi:10.1063/5.0216272.

[18] C. Bannwarth, E. Caldeweyher, S. Ehlert, et al. Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.*, 11:e01493, 2020. doi:10.1002/wcms.1493.

[19] N. Erickson, J. Mueller, A. Shirkov, et al. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020. doi:10.48550/arXiv.2003.06505.

[20] A. Beroual and A. Haddad. Recent advances in the quest for a new insulation gas with a low impact on the environment to replace sulfur hexafluoride (SF_6) gas in high-voltage power network applications. *Energies*, 10(8):1216, 2017. doi:10.3390/en10081216.

[21] H. Hou and B. Wang. Group additivity theoretical model for the prediction of dielectric strengths of the alternative gases to SF_6. *Chemical Journal of Chinese Universities*, 42(12):3709–3715, 2021. doi:10.7503/cjcu20210495.