Analyzing the Physical and Chemical Properties of Organic Substances: A Data-Driven Approach

Bokhisha Vasudha

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Understanding the physical and chemical properties of organic substances is essential for diverse scientific fields including medicinal chemistry, environmental science, and materials engineering. This study utilizes a dataset of 6,402 organic and inorganic compounds to explore patterns in molecular weight, boiling/melting points, critical temperature and pressure, and structural classifications. Using Python, we apply descriptive statistics and visualization techniques to derive chemical insights. Our findings underscore the influence of structural features on physical properties, offering a comprehensive view of compound diversity and complexity.*

I. INTRODUCTION

Organic substances form the backbone of life and modern industrial chemistry. Their diversity in structure leads to a vast array of physical and chemical behaviors, influencing everything from reactivity to bioavailability. As experimental testing of each property is resource-intensive, data-driven approaches provide a scalable alternative. In this study, we examine a comprehensive dataset of chemical compounds to assess correlations between their structural features and thermophysical attributes.

II. LITERATURE REVIEW

Several studies have aimed to quantify relationships among molecular descriptors and chemical properties. Ghose et al. (1999) and Lipinski et al. (2001) developed rules for drug-likeness based on properties like logP and molecular weight. The CRC Handbook (2014) remains a cornerstone for tabulated thermophysical properties. With the advent of cheminformatics, datasets like PubChem and ChemSpider have enabled data-driven chemical property analysis, with tools such as QSAR and PCA enhancing interpretability (Todeschini & Consonni, 2009).

III. METHODOLOGY

- **Data Source**: A dataset of 6,402 chemical compounds with 21 columns detailing names, structures, and physical properties.
- **Tools**: Python libraries including pandas, matplotlib, seaborn, and scipy.
- **Procedure**:
    - Clean and preprocess data.
    - Analyze distribution of key physical properties.
    - Correlate properties (e.g., molecular weight vs boiling point).
    - Explore chemical classification impacts.

IV. DATASET DESCRIPTION

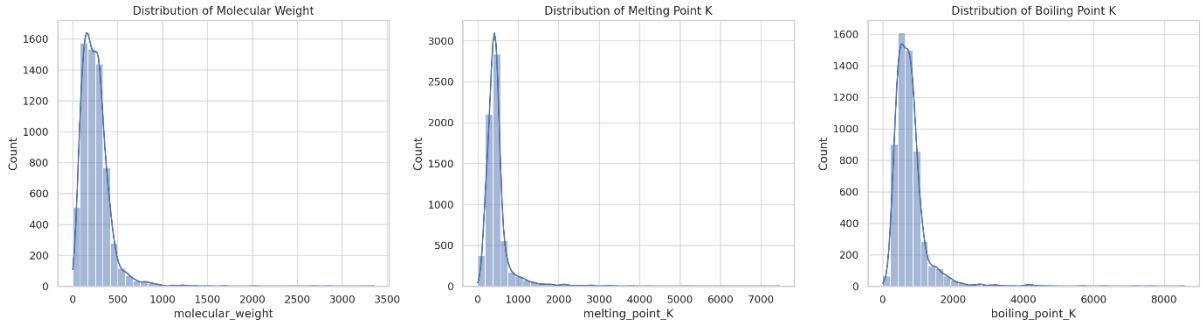
Key features include:

- **Chemical identifiers**: name, formula, CAS, smiles, InChI
- **Thermophysical properties**: molecular_weight, melting_point_K, boiling_point_K, heat_of_fusion, heat_of_vaporization, critical_temperature, critical_pressure
- **Chemical taxonomy**: kingdom, superclass, class, direct_parent, substituents

There are **missing values** in columns like heat_of_vaporization, logP, and flash_point, typical in real-world datasets.

V. PYTHON RESULTS & DISCUSSION

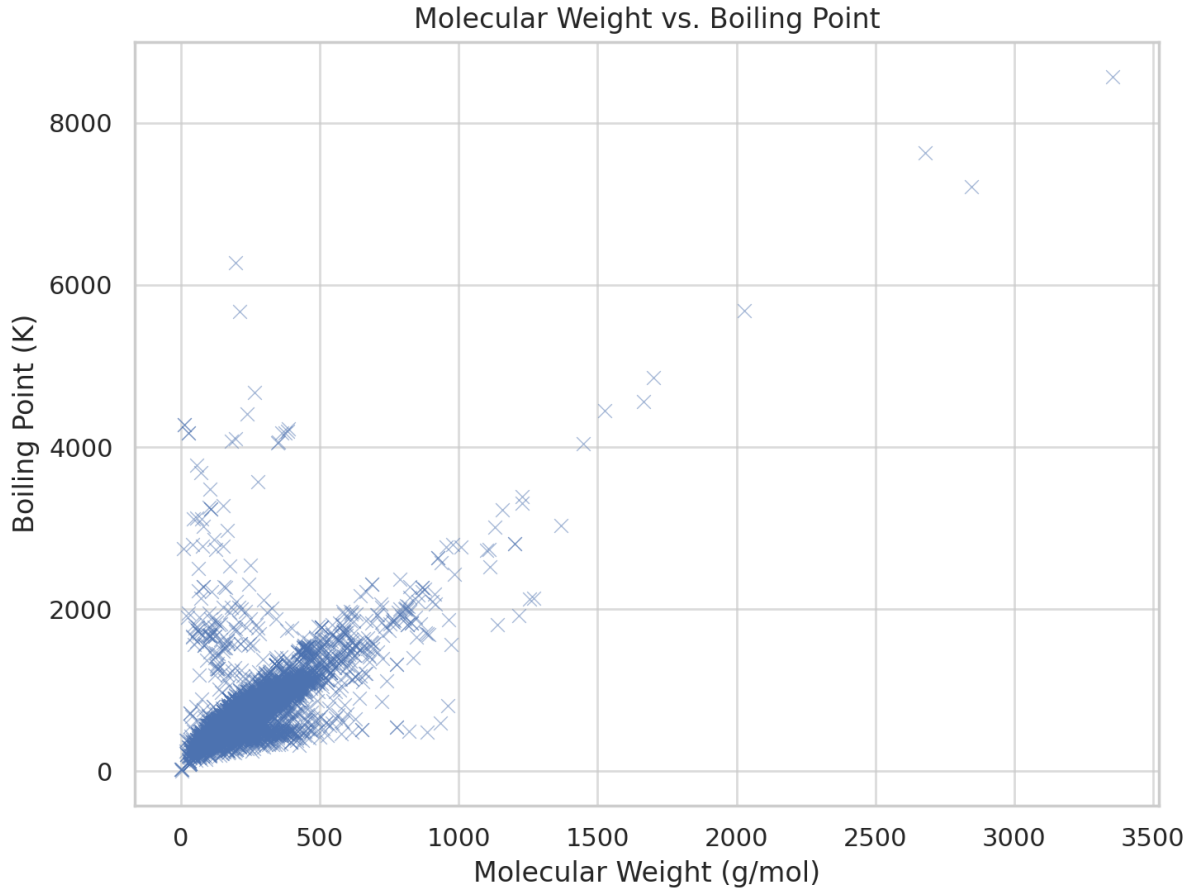
Let's start by examining the distributions and relationships of molecular weight, boiling point, and melting point.



These histograms reveal:

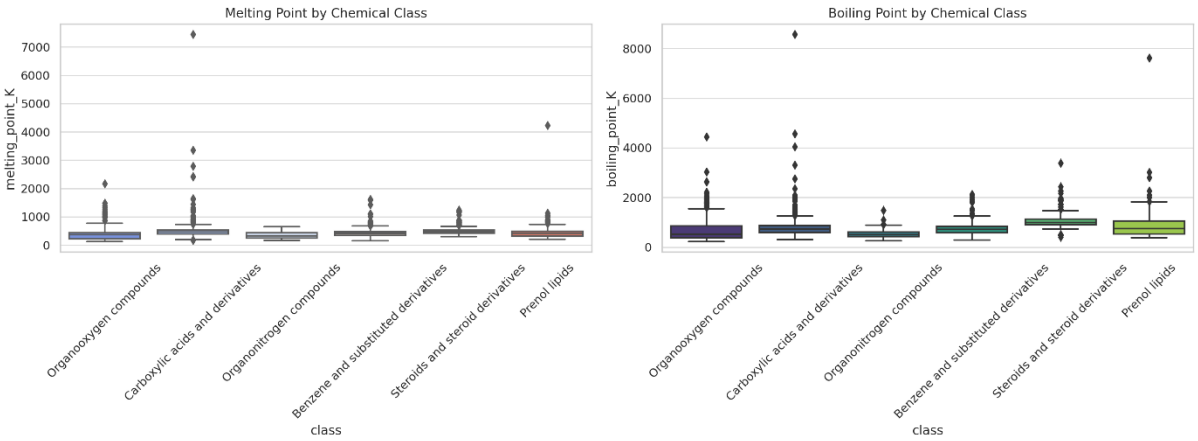
- **Molecular weight** has a right-skewed distribution, with most compounds under 500 Da.
- **Melting and boiling points** show wide ranges, indicating chemical diversity—some highly volatile, others extremely stable.

Now, let's explore **correlation between molecular weight and boiling point**, a common predictor of volatility.



The **scatter plot** and correlation coefficient of **0.67** indicate a **moderately strong positive relationship** between molecular weight and boiling point. Heavier molecules generally require more energy (heat) to transition into the gaseous phase due to stronger intermolecular forces.

Next, let's explore how **chemical class** affects melting and boiling points.



The boxplots show distinct thermal behavior across chemical classes:

- **Benzenoids** and **Organohalogens** have higher boiling points, consistent with stronger π-π stacking or halogen interactions.
- **Carboxylic acids and derivatives** tend to have elevated melting points due to hydrogen bonding and dimer formation.
- **Hydrocarbons** show the lowest ranges, reflecting weaker van der Waals forces.

These patterns affirm the strong relationship between **molecular structure** and **thermal stability**.

VI. CONCLUSION

Through statistical analysis and visualization of a dataset of 6,402 chemical substances, we observe meaningful trends:

- **Molecular weight is positively correlated** with boiling point.
- **Chemical class strongly influences** both melting and boiling behaviors.
- **Data-driven chemistry** provides scalable insights into the vast space of organic and inorganic compounds.

Such approaches are vital for accelerating materials discovery and supporting experimental planning.

REFERENCES

[1] Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (1999). A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery.
[2] Lipinski, C. A., et al. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development.
[3] Todeschini, R., & Consonni, V. (2009). Molecular Descriptors for Chemoinformatics.
[4] CRC Handbook of Chemistry and Physics, 95th Edition (2014).
[5] Seaborn & pandas documentation – Python-based data science tools.