Title

by
Kamal Aslam

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Mathematics

Guelph, Ontario, Canada

ABSTRACT

Kamal Aslam

University of Guelph, 2025

Co-advisors:

Dr. William R. Smith

Dr. Mihai Nica

The accurate prediction of amine pKa values is a significant and persistent challenge within computational chemistry, possessing critical implications for drug discovery and material science. This thesis presents a comprehensive investigation into this problem, systematically evaluating a range of machine learning models and diverse molecular feature representations. The study encompasses Graph Convolutional Networks (GCNs), Multi-Layer Perceptrons (MLPs), Random Forests, Support Vector Regression (SVR), and XGBoost, implemented both as standalone predictive models and within novel fusion architectures. These models were rigorously trained and validated on two distinct datasets of amine compounds: one derived from ChEMBL, filtered to include molecules with up to 12 carbon atoms ("Chembl 12C"), and another from IUPAC sources ("Iupac"). Both datasets were specifically curated to contain amines composed solely of carbon, hydrogen, nitrogen, and oxygen atoms, under standardized experimental conditions.

The central contribution of this research is the demonstration that hybrid modeling approaches, specifically those integrating Graph Neural Networks (GNNs) with established ensemble learning techniques, offer enhanced predictive performance for amine pKa values. Notably, a fusion archi-

tecture combining a GCN with an XGBoost model yielded superior results. Within this optimal configuration, the GCN component effectively utilized a set of 8 aggregated atomic-level descriptors augmented with Gasteiger partial charges ("data G1"). The output from this GCN, serving as learned molecular embeddings, was then processed by an XGBoost model that additionally incorporated Benson group features. The efficacy of this GCN-XGBoost fusion was consistently observed across both datasets and was quantified by improvements in key statistical metrics, including Mean Squared Error (MSE) and the coefficient of determination (R^2).

These findings underscore the potential of synergistic GNN-ensemble models in tackling complex quantitative structure-property relationship (QSPR) tasks. The enhanced accuracy achieved by leveraging graph-based representations for atomic environments, combined with the robust predictive power of gradient boosting methods on curated chemical fragment data, suggests a promising direction for future advancements in pKa prediction and related cheminformatics challenges.

Acknowledgements

All thanks to Dr. Smith, and Dr. Nica

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Predicting Amine pKa using ML Models

This work investigates the prediction of pKa values for amine-containing molecules using a variety of machine learning models and different representations of molecular structure and properties. The study compares the performance of Graph Convolutional Networks (GCNs), Multi-Layer Perceptrons (MLPs), Random Forests, Support Vector Regression (SVR), and XGBoost, both as standalone models and in fusion architectures. The models were trained and evaluated on two distinct datasets: Amines from Chembl up to 12C, and amines from Iupac. Both datasets only contain amines with C, H, N and O atoms.

Dataset

The two datasets used are IuPac and chembl up to 12 carbon amines(Chembl 12C).

Below is a list of filters applied to both IuPac and Chembl 12C

1. Molecule is a primary, secondary or tertiary amine and Is not an Amide(put code here)

2. Molecule contains only C, H, N and O atoms and (remove in final draft: not Sulfur or Silicone)

3. type of measurement is pKaH1, NOT pKaH2 or any other type

4. temp of measurement is 25 celcius

5. There can be no repeat inchi keys. The one with the highest pka value is selected in case there are multiple same inchi keys.

6. If RDKit can't parse the Smiles string, we filter that molecule out

7. take out protonated molecules

IUPAC is used as training and external test set

For filters 1, amines are different from nitrogen containing functional groups (like amides) because of the type of nitrogen atom and its bonding. We're focusing on a specific chemical family and filter 1 ensure our data is relevant to that family, ensuring the model learns patterns unique to amine basicity, rather than mixing the learning with other nitrogen-containing functional groups with distinct pka behaviours.

For filter 2, we're only interested in Carbon capture molecules that contain C,H,N and O. examples of such molecules are amines and Alkanalamines

For filter 3, pkah1 refers to the site that yields the largest pKa reaction. Often, its the most relevant in most studies and has the most data. This ensures consistency and accuracy in what our model is predicting and avoids ambiguity.

For filter 4, pka values are temperature dependent. 25 degrees Celsius has the most available data for amines. By removing the amines at different temperature, we ensure that our model is predicting pka's at 25 degrees.

For filter 5, molecules can appear multiple times in a database due to different experimental conditions. There is no specific reason to select the highest pka other than to keep the method con-

sistent for resolving conflicting entries. And we don't want the model to learn a specific molecule too well. (Make it brief and just say you choose it arbitrarily)

For filter 7, if RDKit can't parse a smiles string, it usually means that the string is malformed or represents and invalid chemical structure. Some of our features are generated directly RDKit. By filtering out these molecules, we ensure the molecules are computationally traceable for descriptor generation and model input.(Include how many RDKit choked on)

For Chembl 12 C, there are 5899 unique molecules and their CX Basic pka is 0.65 - 12.89. Below is the distribution of Amine class and molecular species.

Amine Class	**Percentage (%)**
Primary	35.82
Secondary	34.18
Tertiary	30.01

Table 1.1: Percentage distribution of Amine Class

Molecular Species	**Count**
BASE	3960
NEUTRAL	1928
ACID	11

Table 1.2: Unique values and counts for `Molecular Species`

For Iupac dataset, there are 1530 amines with a range of -10.01 - 14.00. Below are the same graphs for comparison with Chembl.

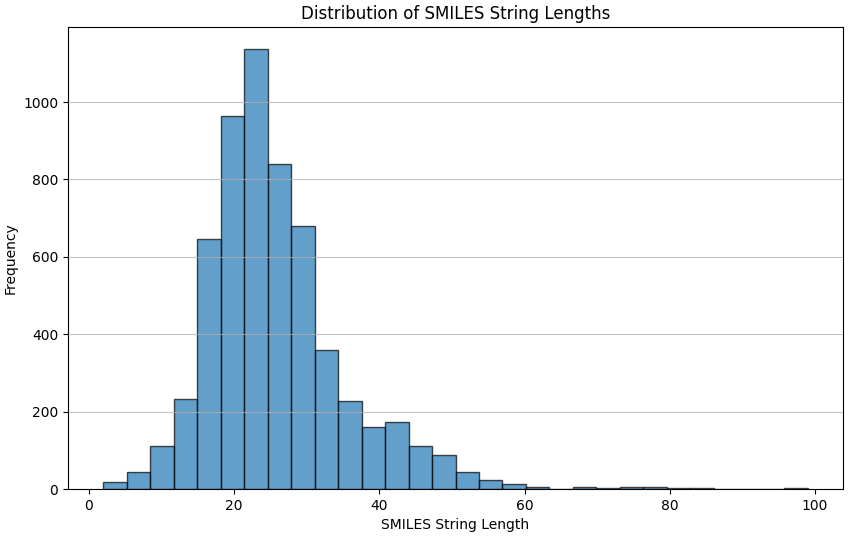

Figure 1.1: Distribution of SMILES String Length in Chembl

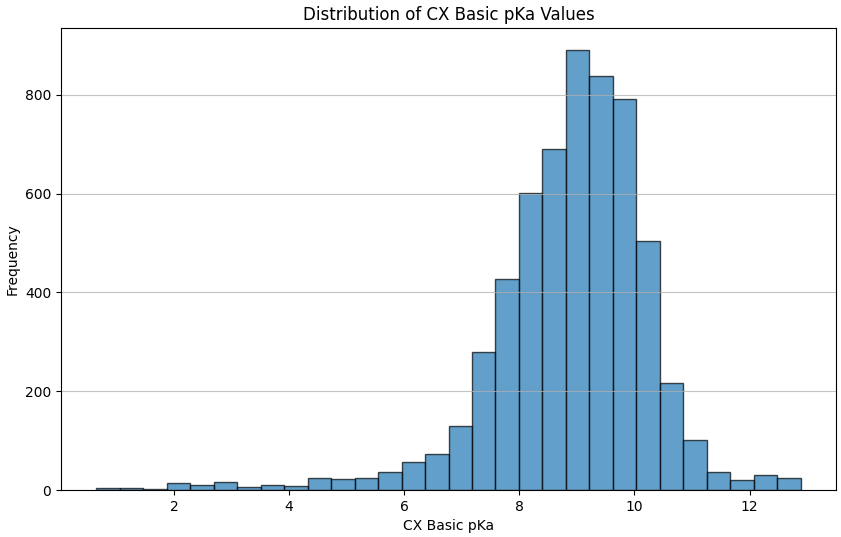

Figure 1.2: Distribution of CX Basic pKa Values in Chembl

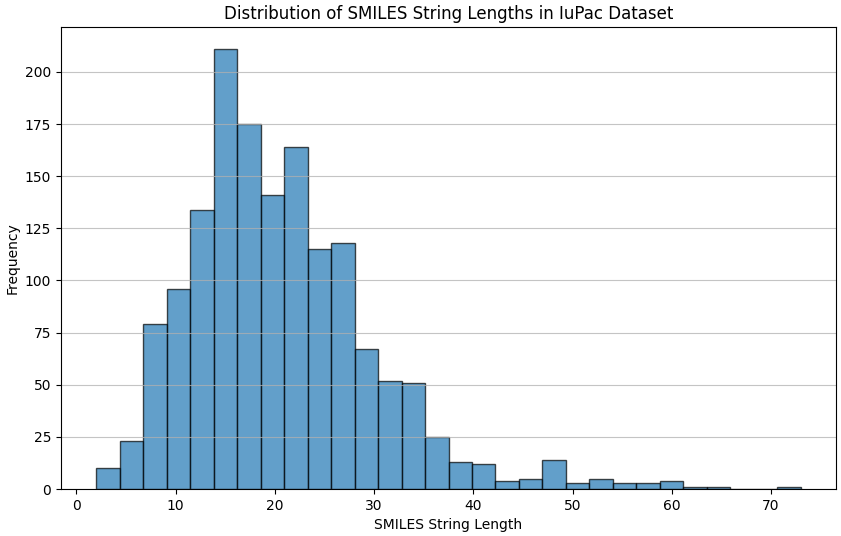

Figure 1.3: Distribution of SMILES String Length in IuPac Dataset

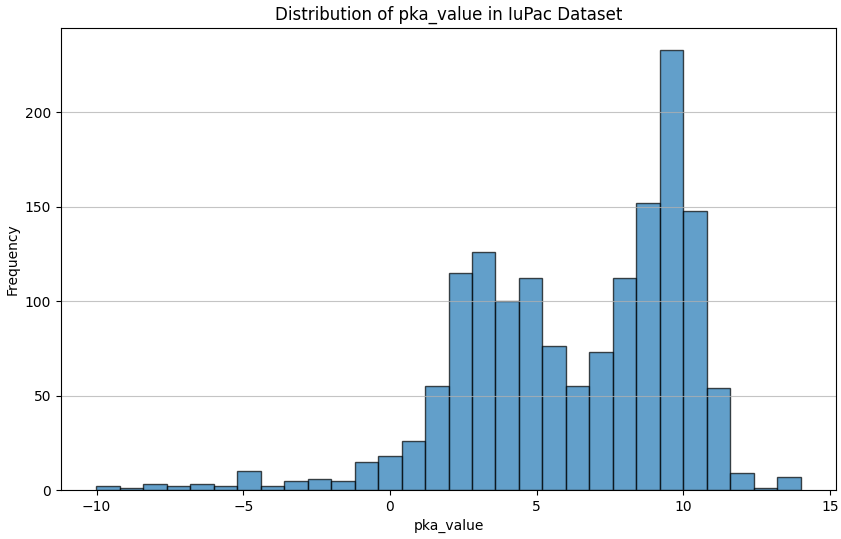

Figure 1.4: Distribution of CX Basic pKa Values in IuPac Dataset

Machine Learning Features

The study utilizes several representations of the amine molecules:

`data_8_atomic_descriptors`: Atomic Descriptors

The ‘data_8_atomic_descriptors‘ dataset employs a set of 8 molecular descriptors, where each molecule is represented by a single 8-dimensional vector. These descriptors are derived from atom-level properties:

1. **Atomic Number**
2. **Formal Charge**
3. **Degree**
4. **Total Number of Hydrogens**
5. **Is Aromatic**
6. **Hybridization**
7. **Is In Ring**
8. **Mass**

It is important to note that for the ‘data_8_atomic_descriptors‘ dataset, these atom-level features are aggregated across all atoms within a molecule to yield a fixed-size molecular descriptor vector, which then serves as the input for machine learning models like GCN.

`data_Gasteiger_Charge`: Atomic Descriptors

The ‘data_Gasteiger_Charge‘ dataset employs a set of 1 molecular descriptors. This feature is used in combination with data_8_atomic_descriptors.

1. **Gasteiger Charge**

`data_AM1BCC_Partial_Charge`: Atomic Descriptors

The ‘data_AM1BCC_Partial_Charge‘ dataset employs a set of 1 molecular descriptors. This feature is used in combination with data_8_atomic_descriptors.

1. **AM1BCC_Partial_Charge**

`data_Computational_Sigma_Profile`: Sigma Profiles

Sigma profiles are descriptors that capture the electronic properties of atoms within a molecule.

`data_Mcginn_Sigma_Profile`: GCN-Predicted Sigma Profiles

The ‘data_Mcginn_Sigma_Profile‘ dataset utilizes sigma profiles that were *predicted* using a Graph Convolutional Network. This allows for the generation of electronic descriptors for a broader set of molecules.

`data_benson_groups`: Benson Groups

The ‘data_benson_groups‘ dataset employs Benson groups, which are structural fragments contributing to a molecule’s thermodynamic properties. These are one-hot encoded and scaled.

`data_Morgan_FingerPrints`: Morgan Fingerprints

The 'data_Morgan_Fingerprints' dataset employs Morgan Fingerprints. Morgan Fingerprints are a type of circular molecular fingerprint commonly used in cheminformatics and machine learning.

Machine Learning Models

The study evaluates the performance of the following machine learning models:

1. **GCN**: A Graph Convolutional Network that directly operates on the molecular graph structure, using the atom-level features described for data_8_atomic_descriptors + data_Gasteiger_Charge (data_G1) and data_8_atomic_descriptors + data_AM1BCC_Partial_Charge(data_A1). The architecture consists of three convolutional layers with batch normalization, ReLU activation, and dropout.

2. **GCN + MLP Fusion**: A model that combines the output of the GCN (trained on data_A1 or data_G1) with a Multi-Layer Perceptron (MLP) trained on the one-hot encoded and scaled Benson group vectors and/or sigma profiles. The outputs of these two branches are concatenated and passed through further linear layers for the final pKa prediction. In total, there are six possibilities for the input features:

   (a) data_A1 + Benson Groups

   (b) data_A1 + Sigma Profile

   (c) data_A1 + Benson Groups + Sigma Profile

   (d) data_G1 + Benson Groups

   (e) data_G1 + Sigma Profile

(f) data_G1 + Benson Groups + Sigma Profile

3. **GCN + XGBoost Fusion**: Similar to the GCN + MLP fusion, but the GCN's output is combined with features derived from Benson groups and/or sigma profile and then inputted into an XGBoost.

4. **MLP**: A Multi-Layer Perceptron trained on different molecular representations, including Sigma Profiles, Benson Groups and Morgan FingerPrints

5. **Random Forest**: An ensemble learning method based on decision trees, applied to Sigma Profiles

6. **SVR**: Support Vector Regression, a powerful method for regression tasks, evaluated on Sigma Profiles

7. **XGBoost**: An optimized gradient boosting algorithm known for its high performance in various machine learning tasks, applied to molecular embeddings extracted from GCN, Sigma Profiles, Benson Groups and Morgan Fingerprints

All models were trained with differing hyperparameters, which were likely tuned to optimize their performance on the respective datasets.

Evaluation

The performance of each model on each dataset was evaluated using several metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R 2)