ChemSolubilityBERTa

Farooq Khan
polymaths-ai
farooq@polymaths.ai

1. Model Overview

ChemSolubilityBERTa is a machine-learning model prototype designed to predict the aqueous solubility of chemical compounds from their SMILES representations. Based on ChemBERTa, a BERT-like transformer-based architecture, ChemBERTa pre-trained on 77M SMILES strings for molecular property prediction. We adapted ChemBERTa to predict solubility values by fine-tuning ChemBERTa with the ESOL (Estimated SOLubility) dataset, a water solubility prediction dataset of 1,128 samples. A user inputs a SMILES string, and the model outputs a log solubility value (log mol/L).

2. Motivation

Accurately predicting aqueous solubility would help overcome molecular engineering bottlenecks in the design of drugs, agrochemicals and materials by accelerating molecular discovery, design and optimization, enabling the virtual screening of thousands of compounds, reducing costly, time-consuming physical experiments.

3. Problem

Predicting solubility is technically challenging because of the number of predictor variables. Each high-level variable's determinant properties drive inherent nonlinear interactions within them. Combined with the nonlinear interactions and interdependencies between variables, this makes predictive modelling challenging.

Physics-based models are constrained by the large number of physicochemical interactions, e.g., intermolecular forces, thermodynamics and quantum effects, which translates into 1) computational expense, the computational power and time needed to model these voluminous interactions, 2) scaling limitations when dealing with complex molecular structures or large datasets and 3) paradoxically insufficient data. While empirical models, e.g., QSPR (Quantitative Structure-Property Relationship) or multiple linear regression (MLR) are constrained by the validity of assumptions in describing how molecules behave in solute.

MLR assumes linear relationships between predictor variables and solubility, and QSPR assumes correlations between molecular structures and pre-identified molecular descriptors to predict solubility. In other words, human assumptions guide empirical

models, approximating relationships seen in the data by fitting observed data to predefined equations, assumptions which are thought sufficient to explain solubility. However, reducing the number of predictors and simplifying assumptions to reduce model complexity doesn't translate into predictive accuracy because nonlinearity and complex interdependent relationships characterise solubility.

Some of the main high-level variables are molecular structure, intermolecular forces, temperature and pH and solid-state properties. Small changes in one variable can drastically alter another variable, affecting solvent-solute interactions, e.g., temperature can affect the molecular structure's conformations.

Similarly, small changes within a variable affect behaviour, e.g., changing functional groups in molecular structure, would inevitably cascade and affect other variables through their interdependent interactions. Nonlinear interconnected dependencies characterise the complexity of predictive aqueous solubility modelling. In other words, the challenge is predicting solubility from predictor variables nearly simultaneously, encapsulating their nonlinear interdependent cascading behaviours, which phenomenologically characterise aqueous solubility.

4. Solution

We construct deep learning neural networks to learn complex patterns from nonlinear data to predict solubility, prediction guided by the data and not by human assumptions. We treat ChemBERTa as a chemical foundation model and fine-tune it to predict aqueous solubility. Fine-tuning ChemBERTa allows the model to retain its ability to generalise molecular property prediction while specialising it to predict solubility values.

5. Methods

5.1 Model architecture

ChemSolubilityBERTa derives from ChemBERTa, based on the BERT (Bidirectional Encoder Representations from Transformers) architecture. ChemBERTa predicts molecular properties from SMILES strings.

5.2 Fine-tuning dataset

We fine-tuned the pre-trained ChemBERTa by training it on the ESOL water solubility dataset of 1,128 compounds with measured solubility values (log mol/L), compounds encoded as SMILES strings. During fine-tuning, the model's weights adapt to the solubility prediction task by minimising the loss between the predicted and actual solubility values. Fine-tuning enables the model to learn solubility nuances while leveraging its molecular understanding gained during pre-training.

5.3 Tokenizer

The ChemBERTa tokenizer transforms the ESOL SMILES strings into numerical representations.

5.4 Training

We trained ChemSolubilityBERTa for 3 epochs with a batch size of 16, 80% training and 20% testing. Fine-tuning involved adjusting the model's internal weights to minimise the difference between the predicted solubility values and the actual solubility values in the dataset.

5.5 Evaluation

Solubility prediction is treated as a regression task, where the model outputs a continuous value representing the predicted solubility (log mol/L). We evaluated the model's prediction accuracy on the test set using mean absolute error (MAE) and R-squared metrics.

6. Results

ChemSolubilityBERTa's initial results are promising. A relatively even spread above and below the regression line suggests the model does not have systematic bias.

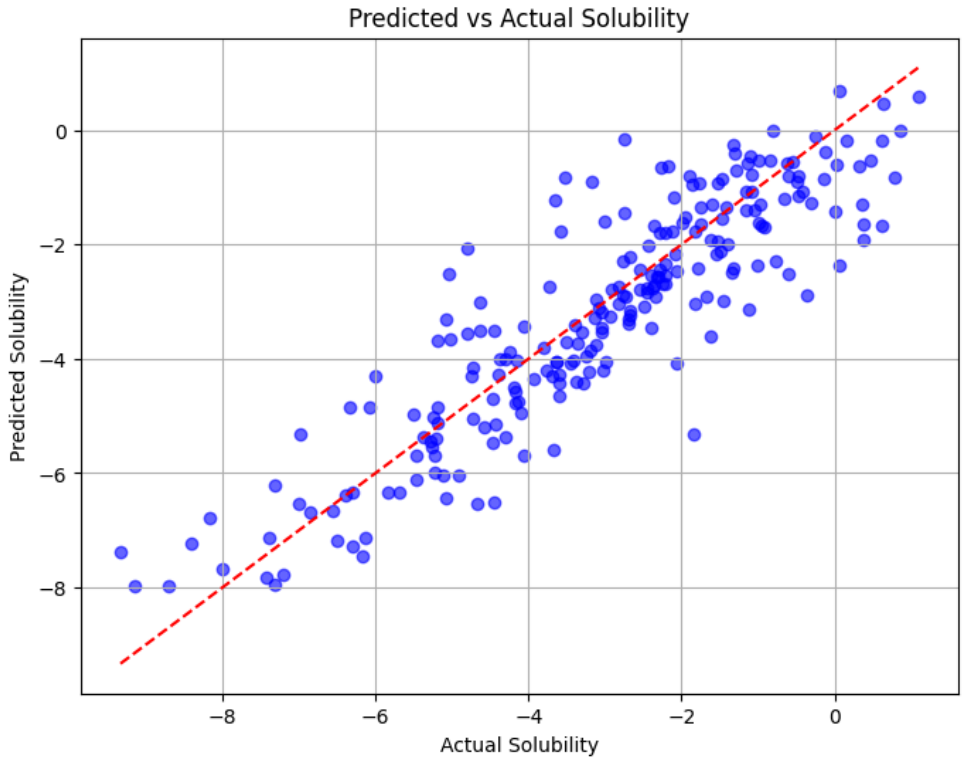

Figure 1: Comparing actual solubility values of compounds from the test dataset with the solubility values predicted by the ChemSolubilityBERTa model. Each blue dot corresponds to a single compound in the test dataset.

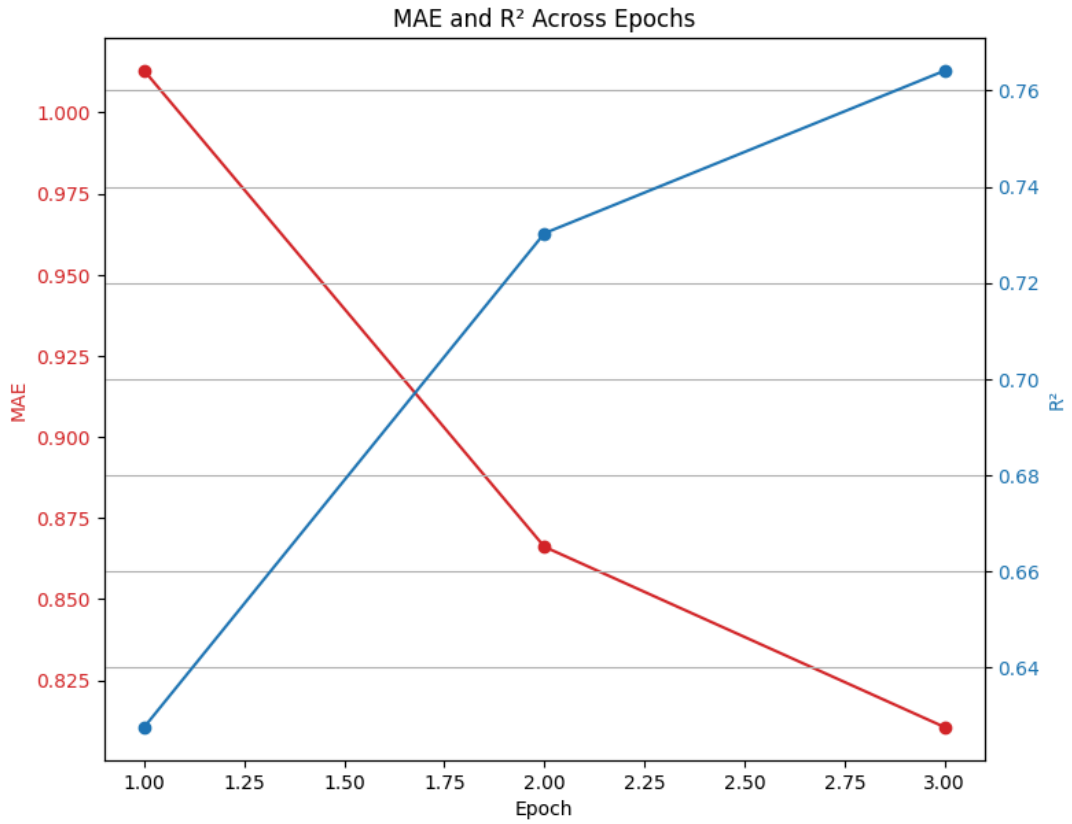

Figure 2: Comparing MAE and R^2 across epochs, plotted on different y-axes.

MAE decreases as training progresses, indicating the model's predictions get closer to the actual solubility values. R^2 increases with training, indicating that the model explains more variance in the data as training progresses. Further work is needed to evaluate potential limitations; predictions may not be as accurate for molecules significantly different from those in the ESOL dataset.

7. Future Work

ChemSolubilityBERTa is a prototype informing our research on AI molecular programming and engineering of molecular complex systems. Future prototyping will incorporate other datasets and expand the ChemSolubilityBERTa model capabilities.

References

Llompart, P., Minoletti, C., Baybekov, S. *et al.* Will we ever be able to accurately predict solubility?. *Sci Data* 11, 303 (2024). https://doi.org/10.1038/s41597-024-03105-6

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, Bharath Ramsundar. ChemBERTa-2: Towards Chemical Foundation Models. ELLIS Machine Learning for Molecule Discovery Workshop, arXiv (2022). https://doi.org/10.48550/arXiv.2209.01712

Seyone Chithrananda, Gabriel Grand, Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. NeurIPS 2020 ML for Molecules Workshop, arXiv (2020). https://doi.org/10.48550/arXiv.2010.09885

Boobier, S., Hose, D.R.J., Blacker, A.J. et al. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. Nat Commun 11, 5753 (2020). https://doi.org/10.1038/s41467-020-19594-z

Sorkun, M.C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. Sci Data 6, 143 (2019). https://doi.org/10.1038/s41597-019-0151-1