Journal, Advances in Mathematical & Computational Sciences
Vol. 12 No. 4, 2024 Series
www.isteams.net/mathematics-computationaljournal

Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA

Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

AIC50: An Innovative Machine Learning Approach to Precise IC50 Prediction

[1]*Obi, E.D., [1]Mbatuegwu, D., [1]Yentumi, J.A. & [2]Ajayi, O.O.
[1] Autogon Inc. 3002 Falls at Fairdale, 77057, USA
[2] Department of Computer Science, Adekunle Ajasin University, Nigeria
E-mail: [1]*eobi@autogon.ai, [1]david@autogon.ai, [1]joshua@autogon.ai, [2] olusola.ajayi@aaua.edu.ng;
Phone: [1]*+1 832 925 1036, [1]+234 816 873 2219, [1]+233 54 100 6410, [2]+27 64 0499906;

ABSTRACT

The half-maximal inhibitory concentration (IC_{50}) is a critical metric in pharmacology, used to evaluate a compound's potency by indicating the concentration needed to inhibit a biological target by fifty percent. Accurate IC_{50} measurement is essential for drug discovery, yet traditional experimental assays are costly and time-consuming, limiting their scalability for early-stage compound screening. Although high-throughput screening (HTS) technologies streamline IC_{50} testing, they remain expensive and resource-intensive. Consequently, computational methods have gained attention as efficient alternatives for predicting IC_{50} values and prioritizing compounds. In this study, we propose an advanced deep learning model for IC_{50} prediction that leverages bidirectional Long Short-Term Memory (LSTM) networks and multi-head attention mechanisms to address the complexity of protein-ligand interactions. Byte Pair Encoding (BPE) was applied to tokenize protein sequences and SMILES strings, providing efficient sub-word representations. Molecular properties such as molecular weight and hydrogen bond characteristics were directly incorporated as features, adding critical biochemical context. Using a comprehensive dataset from the ChEMBL database, containing 94,000 protein-ligand pairs annotated with IC_{50} values, our model achieved a maximum validation accuracy of 83%, outperforming baseline models that lack attention mechanisms. While promising, the model has limitations, including data variability from diverse ChEMBL assay sources and constraints on sequence length, which limits its applicability to large protein sequences and complex ligands. Broader IC_{50} ranges for higher classes also reduce prediction precision. Future improvements include refining IC_{50} class structures, training with multi-source datasets, and increasing model complexity to accommodate longer sequences and structural nuances better. This work advances computational drug discovery by presenting a robust IC_{50} prediction model that integrates sophisticated neural architectures, BPE tokenization, and molecular features, showing potential as an efficient tool for early drug screening. With future enhancements, this model could set a new standard for predictive accuracy, supporting accelerated and cost-effective drug discovery.

Keywords: IC50 prediction, LSTM networks, Computational Drug discovery, Protein-Ligand interactions, Byte Pair Encoding

Obi, E.D., Mbatuegwu, D., Yentumi, J.A. & Ajayi, O.O. (2024): AIC50: An Innovative Machine Learning Approach to Precise IC50 Prediction. Journal of Advances in Mathematical & Computational Science. Vol. 12, No. 4. Pp 31-44
Available online at www.isteams.net/mathematics-computationaljournal. dx.doi.org/10.22624/AIMS/MATHS/V12N4P3

1. INTRODUCTION

1.1 Background

The half-maximal inhibitory concentration (IC_{50}) is the most commonly employed and insightful metric for assessing a drug's effectiveness. It shows the amount of drug required to reduce a biological process by fifty percent, thus offering a measure of the antagonist drug's potency in pharmacological studies [1]. This measure provides insight into a drug's antagonistic capabilities and is essential for determining dosage efficacy in therapeutic contexts [2]. The IC50 value helps researchers understand how much of a compound is needed to achieve a desired biological response, guiding both the selection and optimization of drug candidates in preclinical and clinical studies.

The significance of IC_{50} lies not only in its application to traditional small-molecule drugs but also in its expanding role within modern drug discovery approaches, including biologics, targeted therapies, and multi-target drugs [3,4]. With the shift towards precision medicine, which seeks to tailor treatments to individual patients, understanding the specific inhibitory concentrations of compounds has become increasingly crucial [5,6]. Compounds with lower IC_{50} values are generally regarded as more potent, as they require a smaller concentration to achieve the desired inhibitory effect [6]. This potency is especially relevant when designing drugs that must function within narrow therapeutic windows, where effective doses are close to those that might cause toxicity. Traditionally, determining IC_{50} values relies on experimental assays that measure the inhibitory effects of compounds on particular biological targets, such as enzymes or receptor-ligand interactions [7,8].

These assays can be carried out in vitro, where cells or isolated proteins are exposed to varying concentrations of a compound, or in vivo, where the effects of the compound are measured within a living organism. Although these methods are accurate and reliable, they are inherently resource-intensive [2][9,10]. They often require sophisticated laboratory setups, specialized reagents, and extensive labor, making the high-throughput screening of large compound libraries challenging and costly [11]. In recent years, advancements in high-throughput screening (HTS) technologies have somewhat alleviated these challenges by enabling the simultaneous testing of thousands of compounds [12]. HTS methods use automated workflows and robotics to rapidly assess the inhibitory potential of numerous compounds, streamlining early-stage drug discovery [13]. However, HTS is still an expensive endeavor, with high operating costs and considerable time requirements, especially for experiments that require precise conditions, such as those that mimic physiological environments [14,15]. The rise of computational methods offers a transformative approach to IC_{50} prediction, providing a means to bypass some of the limitations associated with traditional assays and HTS [5][16,17].

Computational techniques such as quantitative structure-activity relationship (QSAR) modeling, molecular docking, and machine learning are increasingly used to predict IC_{50} values based on the chemical and structural properties of compounds [18,19]. QSAR models, for example, seek to correlate the structural characteristics of molecules with their biological activity, allowing for predictive insights into IC_{50} values before any physical compound is synthesized. These models can also capture important molecular features, such as hydrophobicity, electronic distribution, and steric effects, which contribute to binding affinity and, ultimately, inhibitory potency [20].

Machine learning (ML) and deep learning (DL) approaches are further enhancing the predictive power of computational models in drug discovery. Techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer architectures have demonstrated promising results in learning complex representations of molecular structures, enabling accurate IC_{50} predictions [21-23]. By learning from large datasets of protein-ligand interactions, these models can identify intricate patterns in molecular binding and inhibition that may not be immediately apparent through traditional methods. In particular, recurrent models like LSTM (Long Short-Term Memory) networks have been effective for sequential data, such as protein and ligand sequences, while attention-based models excel at highlighting key interactions within the data [24,25].

The utility of computational approaches extends beyond just predicting IC_{50} values; they can also inform the structural optimization of lead compounds. For instance, virtual screening tools can be used to identify promising compounds with low predicted IC_{50} values from large libraries, effectively prioritizing candidates for experimental validation [26-28]. Moreover, by generating interpretable models that highlight important molecular features, computational methods aid in structure-based drug design, where modifications to molecular structure are guided by the model's predicted impact on IC_{50}.

Despite the progress made by computational approaches, IC_{50} prediction remains challenging due to the complexity of biological interactions and the heterogeneity of available data [6]. Protein-ligand interactions are influenced by a myriad of factors, including protein structure, molecular conformation, and environmental conditions such as pH and temperature [7]. Variability in experimental IC_{50} values due to assay conditions and measurement inconsistencies further complicates the development of universally applicable prediction models. Additionally, accurate predictions require large and high-quality datasets with annotated IC_{50} values, but such datasets are often scarce or costly to obtain [29]. As computational methods continue to evolve, integrating diverse data types, such as 3D molecular structures, protein sequences, and biochemical properties, has become a priority. Machine learning models that leverage multi-modal data are better positioned to capture the nuanced interactions between drugs and their targets, improving prediction accuracy [30-32].

For instance, models that combine ligand structural data (e.g., SMILES or 3D conformations) with protein sequence information or molecular dynamics simulations offer a more comprehensive perspective on binding interactions [32-34]. Such multi-input architectures are becoming increasingly common, reflecting the demand for sophisticated, high-accuracy models capable of addressing the complexities inherent in IC_{50} prediction. The development of reliable IC_{50} prediction models has significant implications for drug discovery pipelines. By identifying high-potential candidates earlier and with greater accuracy, these models can streamline the development process, reducing both costs and time to market. This is especially relevant given the high attrition rates in drug discovery, where most compounds fail to make it past initial screening phases. Reliable IC_{50} predictions could enable researchers to focus resources on the most promising compounds, ultimately accelerating the development of effective therapeutics.

1.2 Problem Statement

Despite the advancement of computational models in drug discovery, accurately predicting IC_{50} values remains a challenge. Protein-ligand interactions are inherently complex, influenced by the structural and chemical properties of both the protein and the ligand. Current models often struggle to capture these multifaceted interactions effectively, leading to reduced prediction accuracy. Limitations in model architecture, data representation, and training data quality can result in significant variability in predictions. Moreover, the wide range of IC_{50} values and varying data quality across experimental sources pose additional challenges. There is a clear need for more sophisticated models that can integrate sequence information, molecular properties, and advanced representation techniques to improve prediction outcomes.

For this purpose, we set out to develop a robust neural network model capable of predicting the IC_{50} value ranges for protein-ligand interactions with high accuracy. We designed and implemented an advanced model architecture that leverages bidirectional LSTM networks and multi-head attention mechanisms to capture the intricate dependencies in protein sequences and ligand structures. This architecture Integrated molecular properties of ligands as additional input to enhance predictive performance. Furthermore, we utilized Byte Pair Encoding (BPE) for effective tokenization of both protein sequences and ligand SMILES strings to manage sequence variability. The effectiveness of this approach was then demonstrated through comprehensive training and validation on a large dataset sourced from ChEMBL. This study aims to set a benchmark for future computational models in predicting IC_{50} values, offering a pathway for more precise and interpretable predictions in drug discovery workflows.

2. METHODOLOGY

2.1 Data Collection

We obtained our dataset from the ChEMBL database, a comprehensive repository of bioactive molecules with drug-like properties. The entire ChEMBL dataset was downloaded and imported into a dedicated database server to facilitate efficient data handling. A custom script was developed to iterate through all available proteins in the database, extracting their unique identifiers (IDs) and amino acid sequences. Using these protein IDs, we filtered the assays database table to retrieve associated bioactivity data. In this context, an assay refers to an experimental procedure that measures the activity of a substance (ligand) against a target protein, providing insights into the substance's inhibitory effects.

Each assay entry links to a ligand molecule, allowing us to obtain the corresponding SMILES (Simplified Molecular Input Line Entry System) strings and molecular properties using the molecule's ID. An initial dataset comprising 104,000 entries was generated. To ensure data quality and integrity, entries with missing information were removed, resulting in a final dataset of 94,000 protein-ligand pairs complete with associated molecular properties.

2.2. Data Preprocessing

2.2.1 Tokenization and Encoding

To effectively handle the variability in protein sequences and SMILES strings, we employed Byte Pair Encoding (BPE) for tokenization. BPE is a subword segmentation algorithm that merges the most frequent pairs of bytes in a sequence, which helps in capturing both common and rare patterns in the data. We set the minimum frequency threshold to 10, resulting in vocabulary sizes exceeding 2,000 tokens for both proteins and ligands.

2.2.2 Sequence Padding and Masking

After tokenization, sequences were standardized to fixed lengths to enable batch processing:

- Protein Sequences: Padded to a maximum length of 4,000 tokens.
- SMILES: Padded to a maximum length of 300 tokens.

Sequences shorter than the maximum length were padded with a special token represented by zero to indicate the absence of meaningful data at those positions. This padding ensures uniform input sizes across the dataset. During model training, the embedding layers were configured to mask these zero tokens, effectively ignoring them during computations to prevent them from affecting the learning process.

2.2.3 Ligand Molecular Properties

Ligand properties, specifically molecular weight, hydrogen bond acceptors (HBA), and hydrogen bond donors (HBD), were converted into floating-point numbers. These numerical features were included directly in the model without additional encoding or normalization, allowing the model to leverage this quantitative information during training.

2.3 Model Architecture

Our neural network model was implemented using TensorFlow, a widely used machine learning library. The architecture is designed to process three distinct types of inputs—protein sequences, SMILES strings, and ligand molecular properties—and integrate them for the prediction task.

Input Layers

Tokenized protein sequences are fed into an embedding layer that maps each token to a dense vector representation. The embedding layer is configured to mask zero tokens (padding) to prevent them from influencing the model's learning. Similarly, tokenized SMILES strings are input into an embedding layer with masking of zero tokens. The floating-point ligand properties are input directly into a dense layer to create an initial feature representation.

Embedding and Sequential Modeling

Both the protein and ligand embedding layers transform tokens into dense vectors that capture semantic meanings. The embeddings are passed through bidirectional Long Short-Term Memory (LSTM) layers. These layers process the sequences in both forward and backward directions, capturing contextual information that is essential for understanding the sequential nature of proteins and ligands.

Attention Mechanisms and Pooling

Outputs from the bidirectional LSTMs are fed into multi-head attention layers separately for proteins and ligands. The attention mechanism allows the model to focus on specific parts of the sequences that are most relevant to the interaction, enhancing its ability to identify crucial patterns. The attention outputs undergo both max pooling and average pooling operations. Pooling reduces the dimensionality of the data and aggregates the information learned across the sequences, facilitating more efficient processing in subsequent layers.

Feature Integration and Output

The processed features from the protein branch, ligand branch, and ligand properties branch are concatenated to form a comprehensive feature vector representing the entire protein-ligand interaction. The concatenated features are passed through one or more dense (fully connected) layers. These layers further process the integrated features, enabling the model to learn complex relationships and patterns relevant to the prediction task. The final layer is a dense layer with a softmax activation function. It contains units equal to the predefined IC50 classes, producing a probability distribution over them. Each class corresponds to a specific range on the nanomolar (nM) scale, mapping the classification output to an IC_{50} value range.

2.4 Training Parameters

The training was conducted using four GPUs in a mirrored distribution strategy, allowing parallel processing of data batches. This strategy leverages multiple GPUs to accelerate training by synchronizing the computation of gradients and updates across devices.

2.5 Batch Size and Epochs

A batch size of 1,024 was selected, maximizing the utilization of available GPU memory and computational power. Large batch sizes help in stabilizing gradient estimates and speeding up the training process. The model was trained for 883 epochs. An **epoch** refers to one complete pass through the entire training dataset.

2.6 Learning Rate Schedule and Callbacks

An adaptive learning rate schedule was employed. A callback function monitored the validation loss, reducing the learning rate by 50% whenever a plateau was detected. This approach helps in fine-tuning the model's learning process, allowing it to converge more effectively by taking smaller steps as it approaches a minimum in the loss function. During training, model checkpoints were saved based on the highest achieved accuracy on the validation set. This practice ensures that the best-performing model is retained for evaluation and deployment.

2.7 Validation Strategy

A validation split of 20% was used to partition the dataset into training and validation subsets. The validation set was not used during training but served to evaluate the model's performance on unseen data after each epoch. This strategy provides an unbiased assessment of the model's generalization capability and helps in detecting overfitting.

3. RESULTS

3.1 Performance Metrics

The model achieved a maximum validation accuracy of 83%, indicating that it correctly predicted the IC_{50} class index for 83% of the validation dataset. At the point where this validation accuracy was reached, the corresponding training accuracy was 78%, showing that the model generalizes well to unseen data without significant overfitting. Additionally, the correlation between epoch and accuracy was 0.88, demonstrating a strong positive relationship between increased training epochs and accuracy improvements. The lowest validation loss recorded during training was 0.3884, reflecting the model's ability to minimize prediction errors effectively. This metric is a crucial indicator of the model's performance, as lower loss values suggest better alignment between predictions and actual values.
At the epoch corresponding to the highest validation accuracy, the learning rate had reduced to 2.44e-6, indicating that the adaptive learning rate scheduler played a significant role in fine-tuning the model toward optimal performance.

3.2 Comparative Analysis

The model outperforms baseline models such as standard LSTM networks without attention, which typically achieve validation accuracies of around 75–78%. This demonstrates the impact of incorporating multi-head attention and Bidirectional LSTMs in capturing contextual dependencies within both protein sequences and SMILES strings. The following graphs illustrate the training and validation accuracy trends over the 883 epochs, showing the steady improvement in performance as training progresses. Notably, validation accuracy closely follows the training accuracy, indicating the model's ability to generalize well without overfitting.

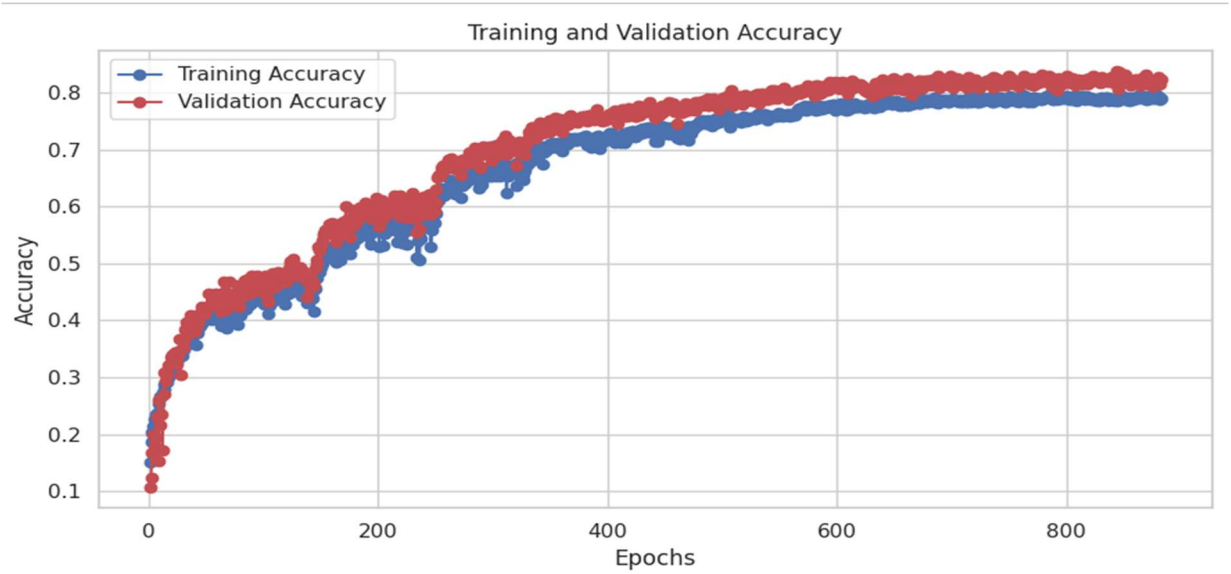

Fig. 1: Training and Validation Accuracy trend over 883 epochs

Journal, Advances in Mathematical & Computational Sciences
Vol. 12 No. 4, 2024 Series
www.isteams.net/mathematics-computationaljournal

Additionally, the training and validation loss curves highlight the model's convergence. Both losses decreased steadily, with validation loss reaching a value of 0.48 near the end of training, aligning with the point of highest validation accuracy.

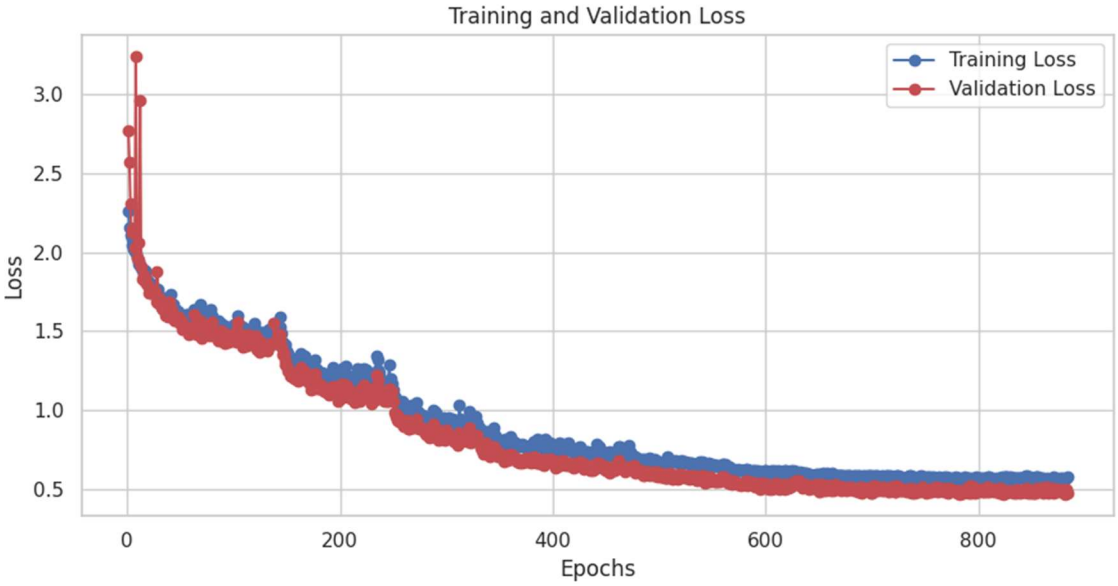

Fig. 2: Training and Validation Loss over 883 epochs

The learning rate schedule shows how the learning rate was dynamically adjusted over the training process. As validation loss plateaued, the learning rate decreased, with 2.44e-6 being the value at the point of maximum validation accuracy. This adaptive learning strategy was instrumental in achieving optimal performance.

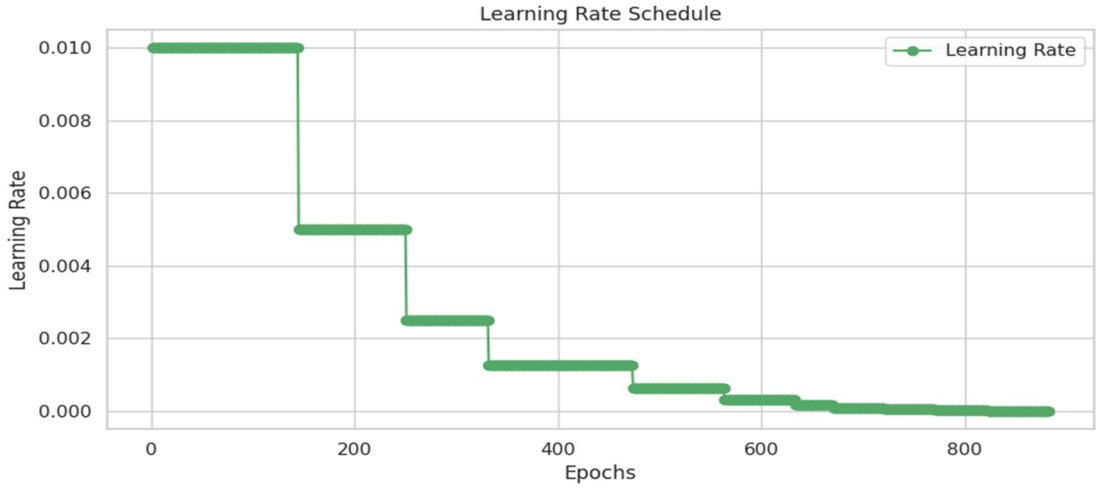

Fig. 3: Learning Rate Adjustment over 883 epochs

4. DISCUSSION

4.1 Interpretation of Results and Their Significance

The results of this study demonstrate the effectiveness of the proposed neural network model for predicting IC_{50} values for protein-ligand interactions. Achieving a maximum validation accuracy of 83% indicates that the model is capable of correctly classifying the IC_{50} value range for most cases in the validation set. This is significant in the context of drug discovery, as accurately predicting the IC_{50} range can streamline the process of identifying promising compounds for further experimental validation. The correlation coefficient of 0.91 between epoch and accuracy supports the reliability of the training process, indicating that the model's performance consistently improved over time. The low validation loss (0.3884) further underscores the model's ability to generalize to unseen data.

4.2 Contribution of the Model's Architecture

The architecture played a crucial role in the model's success. The combination of bidirectional LSTMs allowed the model to capture contextual information from both directions of the protein sequences and SMILES strings, ensuring a more comprehensive understanding of sequence dependencies. The multi-head attention layers enhanced the model's ability to focus on important parts of the input, helping it discern critical patterns within the sequences that contribute to protein-ligand interactions. Additionally, incorporating ligand molecular properties as direct numerical inputs provided supplementary biochemical context that improved the model's predictions.

The use of Byte Pair Encoding (BPE) for tokenizing the protein sequences and SMILES strings resulted in efficient representation and processing of the input data, helping the model learn from subword information. The adaptive learning rate strategy and use of a mirrored distribution strategy across GPUs were pivotal in stabilizing the training process and enhancing convergence.

4.3 Limitations

Despite the promising results, this study faced several limitations:

1. **Experimental Nature of the ChEMBL Database**: The model's predictions are dependent on the reliability of the ChEMBL database entries. Since ChEMBL data is sourced from various experimental assays with potential variability in measurement accuracy, any inaccuracies or inconsistencies in the database may impact the model's training and evaluation.
2. **Protein Sequence Lengths**: The maximum length for protein sequences was capped at 4,000 tokens, which poses a limitation for handling exceptionally large protein sequences. Sequences longer than this threshold were not represented in the training data, potentially limiting the model's applicability to more complex proteins with greater lengths.
3. **SMILES String Lengths**: The maximum SMILES string size of 300 tokens could restrict the model's performance for larger, more structurally complex ligands. This constraint might limit the model's effectiveness when predicting larger molecule interactions.
4. **Higher Predicted Range Classes**: The class structure for IC_{50} prediction inherently involves larger ranges for higher IC_{50} values. This results in broader sample spacing for those classes, making it more difficult to predict an exact IC_{50} value as the class index increases. This could affect the precision of predictions in the higher IC_{50} ranges, potentially leading to less granular outcomes in these cases.

4.4 Potential Improvements and Future Research Directions

To further enhance the model's performance and overcome current limitations, several improvements and future research directions can be considered:

1. **Experimenting with the Number of Classes**: Adjusting the number of classes for IC_{50} prediction could provide more refined predictions. Increasing the number of classes could reduce the sample spacing within higher ranges, potentially improving the model's precision for predicting higher IC_{50} values.
2. **Training with Combined Datasets**: Integrating additional datasets from sources beyond ChEMBL could enrich the training data and make the model more robust. By incorporating data from multiple databases, the model could be exposed to a broader variety of protein-ligand interactions, improving generalization.
3. **Increasing Model Complexity**: Introducing more sophisticated architectural components, such as deeper LSTM layers or additional attention mechanisms, could help the model capture more complex interactions between proteins and ligands. Exploring transformer-based architectures or hybrid models that combine CNNs with LSTMs and attention layers may yield better performance.
4. **Handling Longer Sequences**: Developing methods for segmenting or compressing longer protein sequences and SMILES structures into manageable input sizes could help the model deal with more extensive sequences that exceed current input limits.

By addressing these limitations and pursuing these potential improvements, future research can continue to advance the field of IC_{50} prediction, making computational models more accurate and practical for use in drug discovery and development.

5. CONCLUSION

This study successfully demonstrated that an integrated deep learning approach using bidirectional LSTMs, multi-head attention, and ligand property inputs can effectively predict IC_{50} value ranges for protein-ligand interactions, achieving a maximum validation accuracy of 83%. The results indicate that this approach can serve as a powerful tool for preliminary drug screening by offering reliable, interpretable predictions that align well with real-world data.

5.1 Contributions of the Study

This study contributes to the field of computational drug discovery by proposing a novel approach that integrates several advanced techniques to predict IC_{50} values more accurately. Key contributions include:

- **Innovative Architecture**: The use of **bidirectional LSTMs** and **multi-head attention** mechanisms for protein and ligand sequence analysis, which improves the model's ability to capture complex, context-dependent interactions.

- **Enhanced Data Representation**: The application of **BPE tokenization** for protein sequences and SMILES strings, allowing for more efficient and informative encoding of biochemical data.
- **Molecular Property Integration**: Incorporation of ligand-specific properties, such as molecular weight and hydrogen bond characteristics, to provide additional context for the model.
- **Robust Training Strategy**: Leveraging a large dataset from **ChEMBL** and employing adaptive learning strategies to achieve significant validation accuracy and generalizability.

5.2 Potential Impact and Future Outlook

This work's potential impact lies in its ability to accelerate drug discovery by streamlining the initial evaluation of compound efficacy. By incorporating suggested improvements—such as more detailed IC_{50} class structures and the use of expanded training datasets—the model could become an even more valuable asset. Future research may build on this foundation, exploring more complex architectures and techniques to expand the scope of predictive capabilities in biochemical modeling, ultimately aiding researchers in making faster, data-driven decisions in drug development.

REFERENCES

1. Aykul, S., & Martinez-Hackert, E. (2016). Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. Analytical Biochemistry, 508, 97-103. https://doi.org/10.1016/j.ab.2016.06.025
2. Bag, A., & Ghorai, P. K. (2016). Development of quantum chemical method to calculate half maximal inhibitory concentration (IC50). Molecular Informatics, 35(5), 199-206. https://doi.org/10.1002/minf.201501004
3. Batool, M., Ahmad, B., & Choi, S. (2019). A structure-based drug discovery paradigm. International Journal of Molecular Sciences, 20(11), Article 2783. https://doi.org/10.3390/ijms20112783
4. Blay, V., Tolani, B., Ho, S. P., & Arkin, M. R. (2020). High-throughput screening: Today's biochemical and cell-based approaches. Drug Discovery Today, 25(10), 1807-1821. https://doi.org/10.1016/j.drudis.2020.07.024
5. Bomane, A., Gonçalves, A., & Ballester, P. J. (2019). Paclitaxel response can be predicted with interpretable multi-variate classifiers exploiting DNA-methylation and miRNA data. Frontiers in Genetics, 10, Article 1041. https://doi.org/10.3389/fgene.2019.01041
6. Cadow, J., Born, J., Manica, M., Oskooei, A., & Rodríguez Martínez, M. (2020). PaccMann: A web service for interpretable anticancer compound sensitivity prediction. Nucleic Acids Research, 48(W1), W502-W508. https://doi.org/10.1093/nar/gkaa327
7. Caldwell, G. W., Yan, Z., Lang, W., & Masucci, J. A. (2012). The IC50 concept revisited. Current Topics in Medicinal Chemistry, 12.

8. Chang, Y., Hawkins, B. A., Du, J. J., Groundwater, P. W., Hibbs, D. E., & Lai, F. (2023). A guide to in silico drug design. Pharmaceutics, 15(1), Article 49. https://doi.org/10.3390/pharmaceutics15010049
9. Cho, K., Choi, E. S., Kim, J. H., Son, J. W., & Kim, E. (2022). Numerical learning of deep features from drug-exposed cell images to calculate IC50 without staining. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-10643-9
10. Cokelaer, T., Chen, E., Iorio, F., Menden, M. P., Lightfoot, H., Saez-Rodriguez, J., & Garnett, M. J. (2018). GDSCTools for mining pharmacogenomic interactions in cancer. Bioinformatics, 34, 1226-1228.
11. Delre, P., Lavado, G. J., Lamanna, G., Saviano, M., Roncaglioni, A., Benfenati, E., Mangiatordi, G. F., & Gadaleta, D. (2022). Ligand-based prediction of hERG-mediated cardiotoxicity based on the integration of different machine learning techniques. Frontiers in Pharmacology, 13. https://doi.org/10.3389/fphar.2022.951083
12. Filipp, F. V. (2017). Precision medicine driven by cancer systems biology. Cancer Metastasis Reviews, 36, 91-108.
13. Ghasemi, M., Nabipour, I., Omrani, A., Alipour, Z., & Assadi, M. (2016). Precision medicine and molecular imaging: New targeted approaches toward cancer therapeutic and diagnosis. American Journal of Nuclear Medicine and Molecular Imaging, 6, 310-327.
14. Gomis-Tena, J., Brown, B. M., Cano, J., Trenor, B., Yang, P. C., Saiz, J., Clancy, C. E., & Romero, L. (2020). When does the IC50 accurately assess the blocking potency of a drug? Journal of Chemical Information and Modeling, 60(3), 1779-1790. https://doi.org/10.1021/acs.jcim.9b01085
15. Guvenc, B., Mamitsuka, H., & Kaski, S. (2019). Improving drug response prediction by integrating multiple data sources: Matrix factorization, kernel, and network-based approaches. Briefings in Bioinformatics, 22(1), 346-359. https://doi.org/10.1093/bib/bbz153
16. Imran, A., Posokhova, I., Qureshi, H. N., Masood, U., Riaz, M. S., Ali, K., John, C. N., Hussain, M. I., & Nabeel, M. (2020). AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Informatics in Medicine Unlocked, 20, Article 100378.
17. Joo, M., Park, A., Kim, K., Son, W. J., Lee, H. S., Lim, G., Lee, J., Lee, D. H., Ahn, J., Kim, J. H., Ahn, T., & Nam, S. (2019). A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. International Journal of Molecular Sciences, 20(24). https://doi.org/10.3390/ijms20246276
18. Koras, K., Juraeva, D., Kreis, J., Mazur, J., Staub, E., & Szczurek, E. (2020). Feature selection strategies for drug sensitivity prediction. Scientific Reports, 10(1), Article 9377. https://doi.org/10.1038/s41598-020-65927-9
19. Kulkarni, S., Seneviratne, N., Baig, M. S., & Khan, A. H. A. (2020). Artificial intelligence in medicine: Where are we now? Academic Radiology, 27, 62-70.

20. Kurilov, R., Haibe-Kains, B., & Brors, B. (2020). Assessment of modelling strategies for drug response prediction in cell lines and xenografts. Scientific Reports, 10(1), Article 2849. https://doi.org/10.1038/s41598-020-59656-2
21. Lee, Y., & Nam, S. (2021). Performance comparisons of alexnet and googlenet in cell growth inhibition IC50 prediction. International Journal of Molecular Sciences, 22(14). https://doi.org/10.3390/ijms22147721
22. Liu, P., Li, H., Li, S., & Leung, K. S. (2019). Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. BMC Bioinformatics, 20(1), Article 408. https://doi.org/10.1186/s12859-019-2910-6
23. Ma, M. S., Sundaram, S., Lou, L., Agarwal, A., Chen, C. S., & Bifano, T. G. (2023). High throughput screening system for engineered cardiac tissues. Frontiers in Bioengineering and Biotechnology, 11. https://doi.org/10.3389/fbioe.2023.1177688
24. Musa, A., Ghoraie, L. S., Zhang, S. D., Glazko, G., Yli-Harja, O., Dehmer, M., Haibe-Kains, B., & Emmert-Streib, F. (2017). A review of connectivity map and computational approaches in pharmacogenomics. Briefings in Bioinformatics, 18, 903.
25. Musulin, J., Baressi Šegota, S., Štifanić, D., Lorencin, I., Andelić, N., Šušteršič, T., Blagojević, A., Filipović, N., Cabov, T., & Markova-Car, E. (2021). Application of artificial intelligence-based regression methods in the problem of COVID-19 spread prediction: A systematic review. International Journal of Environmental Research and Public Health, 18, Article 4287.
26. Parca, L., Pepe, G., Pietrosanto, M., Galvan, G., Galli, L., Palmeri, A., Sciandrone, M., Ferrè, F., Ausiello, G., & Helmer-Citterich, M. (2019). Modeling cancer drug response through drug-specific informative genes. Scientific Reports, 9(1), Article 15222. https://doi.org/10.1038/s41598-019-50720-0
27. Park, A., Joo, M., Kim, K., Son, W. J., Lim, G. T., Lee, J., Kim, J. H., Lee, D. H., & Nam, S. (2022). A comprehensive evaluation of regression-based drug responsiveness prediction models, using cell viability inhibitory concentrations (IC50 values). Bioinformatics, 38(10), 2810-2817. https://doi.org/10.1093/bioinformatics/btac177
28. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., & Ester, M. (2019). MOLI: Multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics, 35(14), i501-i509. https://doi.org/10.1093/bioinformatics/btz318
29. Shryock, R. H. (2017). The development of modern medicine: An interpretation of the social and scientific factors involved. University of Pennsylvania Press.
30. Stanfield, Z., Coskun, M., & Koyuturk, M. (2017). Drug response prediction as a link prediction problem. Scientific Reports, 7, Article 40321.
31. Wang, M. (2023). High-throughput screening technologies for drug discovery. Theoretical and Natural Science, 15(1), 18-23. https://doi.org/10.54254/2753-8818/15/20240387
32. Wang, S., & Li, J. (2020). Modular within and between score for drug response prediction in cancer cell lines. Molecular Omics, 16, 31-38.

33. Wei, D., Liu, C., Zheng, X., & Li, Y. (2019). Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. BMC Bioinformatics, 20(1), Article 44. https://doi.org/10.1186/s12859-019-2608-9
34. Zhang, C., Xiang, J., Xie, Q., Zhao, J., Zhang, H., Huang, E., Shaw, P., Liu, X., & Hu, C. (2021). Identification of influenza PAN endonuclease inhibitors via 3D-QSAR modeling and docking-based virtual screening. Molecules, 26(23). https://doi.org/10.3390/molecules26237129