Enhancing AMPL Drug Discovery Predictions: Model Selection, Descriptor Performance, and Uncertainty Quantification

Chongye Feng

April 24, 2024

Organization: Accelerating Therapeutics for Opportunities in Medicine (ATOM) Consortium – Open Data and Models Group
Mentors: Ya Ju Fan, PhD and Amanda Paulson, PhD
MSSE Mentor: Brandon Allgood, PhD
Collaborators: Radhika Sahai (MSSE student), Jaycee Pang (MSSE student)
Topics: machine learning, classification, uncertainty quantification, drug discovery

Abstract

Machine learning, applied to the drug discovery pipeline, holds promise for streamlining the process of identifying potential drug candidates. This project aimed to build and compare various classification models to determine their efficacy in classifying NimA-related kinase binders and inhibitors, with a particular focus on the performance of different descriptors and sampling methods. Our comparative analysis revealed that Morgan Fingerprints (MF) descriptors outperform Molecular Operating Environment (MOE) descriptors in terms of recall scores, critical for identifying active compounds. Additionally, while Oversampling (OS) typically yielded higher accuracy, Undersampling (US) proved more effective for binding datasets in enhancing recall. Gaussian Process (GP) models generally offered better reliability, except in cases using MOE descriptors where Random Forest (RF) with US was more suited for the binding dataset. These findings underscore the importance of tailored model and method selection in drug discovery, particularly when handling imbalanced datasets. The study's insights into descriptor and model performance are pivotal for optimizing the identification process of potential drug candidates, illuminating future paths for enhancing model accuracy and reliability.

1 Introduction

The drug discovery process is time and resource intensive, often taking 10-15 years and costing billions of dollars to bring a single new drug to market [2]. Beyond the extensive clinical research phase, one significant bottleneck is the discovery and optimization of potential hit compounds. This phase involves screening, evaluating, and optimizing libraries containing millions of molecules, a task that requires substantial time and resources. Leveraging machine learning (ML) in this process presents a promising avenue to enhance efficiency and effectiveness.

Machine learning models, including Gaussian Processes (GP) and Random Forest (RF) classifiers, can significantly expedite the identification and optimization of potential drug candidates, providing a faster and more cost-effective alternative to traditional methods. However, the performance of these models is heavily dependent on the quality of the training data. Datasets in this domain are often imbalanced, featuring a disparity between the number of active (positive) and inactive (negative) compounds, which can hinder the model's ability to learn and generalize effectively. Methods like Oversampling (OS) and Undersampling (US) are employed to address these imbalances, with each technique having its implications on the model performance.

This paper discusses the implementation and evaluation of various machine learning classification models within the context of the Accelerating Therapeutics for Opportunities in Medicine (ATOM) Consortium's efforts to improve the drug discovery pipeline. Specifically, we focus on the classification of NimA-related kinase (NEK) binders and inhibitors, analyzing different models, descriptor types such as Molecular Operating Environment (MOE) and Morgan Fingerprints (MF), sampling techniques, and the role of uncertainty quantification in enhancing model reliability and performance. Our study contributes to the ongoing discourse on the application of machine learning in drug discovery, aiming to identify methodologies that can predict a molecule's binding or inhibition activity with higher accuracy and reliability.

2 Background

The integration of machine learning (ML) in drug discovery has transformed traditional methodologies, offering novel pathways to accelerate and refine the identification and optimization of therapeutic compounds. Central to this advancement is the concept of uncertainty quantification in ML, particularly through Gaussian Processes (GPs) and Random Forest (RF) classifiers, which are instrumental in handling the inherent complexities of biological data.

2.1 Gaussian Processes and Random Forest Classifiers

Gaussian Processes (GPs) are non-parametric, probabilistic models that provide a versatile framework for regression and classification tasks. They model uncertainty by defining a distribution over possible functions that could fit the data, characterized by mean and covariance functions. These models are particularly valuable in drug discovery, where data may be sparse and imbalanced, as they can dynamically update their predictions based on new evidence, offering a robust way to handle uncertainty [7].

Random Forest (RF) classifiers operate through an ensemble of decision trees to make predictions. Each tree in the forest contributes to the final output through a majority voting mechanism, helping the model to capture diverse patterns in the data and provide robustness against noise. The ensemble nature of RF also facilitates a natural quantification of uncertainty, as the variance in predictions across the trees can indicate the confidence level of the model's outputs [15].

2.2 Descriptors and Sampling Methods

Descriptors play a crucial role in transforming chemical information into a format that machine learning models can process. Molecular Operating Environment (MOE) descriptors capture physicochemical properties, while Morgan Fingerprints (MF), a type of circular fingerprint, encode the presence or absence of substructures [21]. The choice between these descriptors can significantly affect model performance, particularly in terms of recall and precision.

Sampling methods such as Oversampling (OS) and Undersampling (US) are techniques used to correct imbalances in training datasets. Oversampling increases the frequency of the minority class by duplicating instances or generating synthetic instances, while undersampling reduces the frequency of the majority class. These methods impact not only the balance of the dataset but also the model's ability to generalize from the training data to unseen data.

3 Testing, Experimentation, Verification and Validation

The raw datasets used in this study detail the binding and inhibition characteristics of molecules for NEK9, labeled with binary target values of 0 or 1, indicating inactivity or activity, respectively. A rigorous strategy for testing and validating our machine learning models was developed, incorporating detailed data preparation, model training, and a comprehensive validation process.

3.1 Data Preparation

The datasets were prepared using two types of descriptors:

- **Molecular Operating Environment (MOE)** descriptors, derived from the compounds' SMILES strings, capture a wide range of physicochemical, topological, and molecular properties. These descriptors were scaled using `StandardScaler()` to normalize the data, improving the input quality for machine learning models.

- **Morgan Fingerprints (MF)** were generated using the Chem library from RDKit, with a radius of 2 and a bit length of 2048. This method encodes the presence of substructural features, providing a binary representation of molecular structure that is highly effective for pattern recognition.

Given the imbalanced nature of the datasets, we employed both oversampling and undersampling techniques to achieve a more balanced class distribution:

- **Oversampling** was performed using SMOTE (Synthetic Minority Oversampling Technique), which synthesizes new examples in the feature space.

- **Undersampling** was implemented using RandomUnderSampler to reduce the size of the majority class by randomly selecting a subset of data for the non-active compounds.

3.2 Model Training and Validation

Gaussian Process (GP) and Random Forest (RF) classifiers were chosen based on their ability to incorporate measures of uncertainty, which is crucial for reliable decision-making in drug discovery. The training process involved:

- Dividing the dataset into five folds to ensure each model was trained and validated on a representative sample of the data.

- Employing cross-validation to assess model performance and generalizability, thereby ensuring the robustness and reliability of the predictions.

3.3 Experimentation with Sampling Techniques

The choice of sampling techniques was critical to address dataset imbalances and to enhance model performance. Experimentation with SMOTE and RandomUnderSampler helped to fine-tune the models' abilities to detect less prevalent, but critical, active compounds. The impact of these techniques on model learning and generalization was rigorously analyzed.

3.4 Verification and Validation

The models underwent a stringent verification and validation process to meet pre-defined accuracy, precision, recall, and specificity criteria. This involved:

- A series of iterative tests and refinements.

- Evaluation against a reserved test set to confirm the models' efficacy in predicting the activity of NEK9 inhibitors and binders accurately.

This rigorous validation is crucial in determining the practical applicability of the models in real-world drug discovery contexts, where prediction accuracy significantly impacts research and development directions.

In conclusion, our meticulous approach to testing, experimentation, verification, and validation underscores the rigorous standards required to develop reliable and effective machine learning models for drug discovery, specifically in identifying potential inhibitors and binders for NEK9. For more details on the source code and implementation, refer to the Appendix section of this report.

4 Results

This section presents the results of our machine learning models, focusing on recall and accuracy scores as they relate to different descriptors and model configurations. The results are visualized through a series of plots that illustrate the performance of each model under various conditions.

4.1 Overall Performance by Model and Descriptor

Figures 1 and 2 show the recall and accuracy scores by model and descriptor for the binding and inhibition datasets, respectively. These figures provide an overview of how each combination of model and descriptor performs, highlighting the optimal configurations for each scenario.

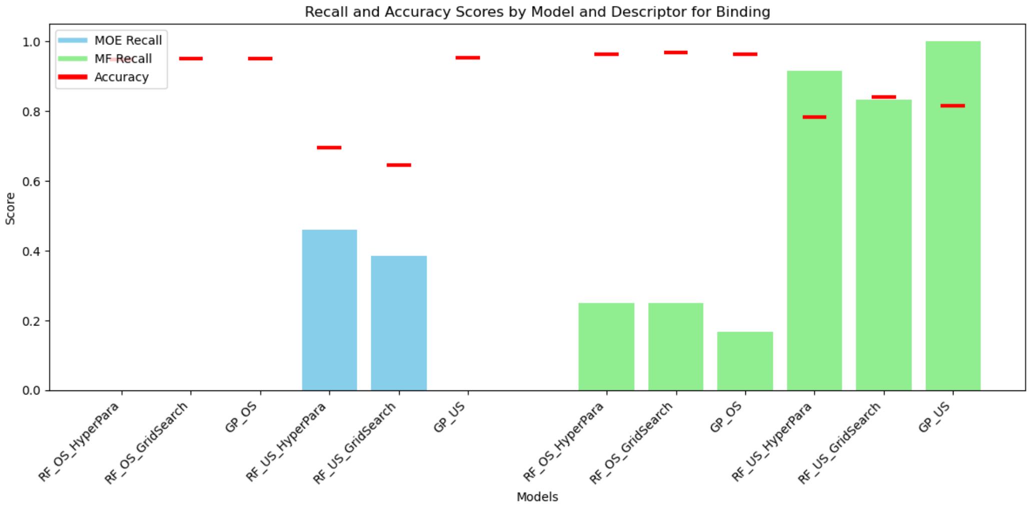


Figure 1: Recall and accuracy scores by model and descriptor for the binding dataset.

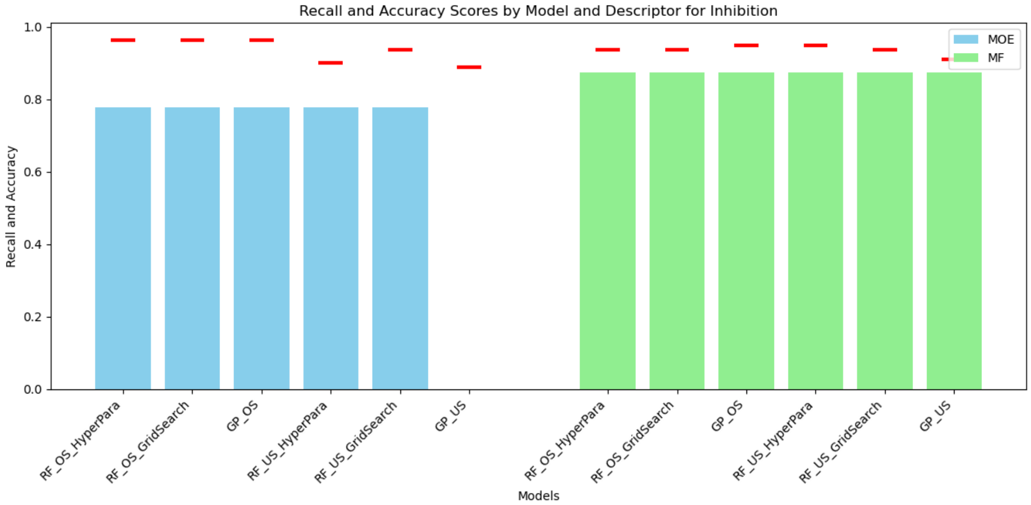
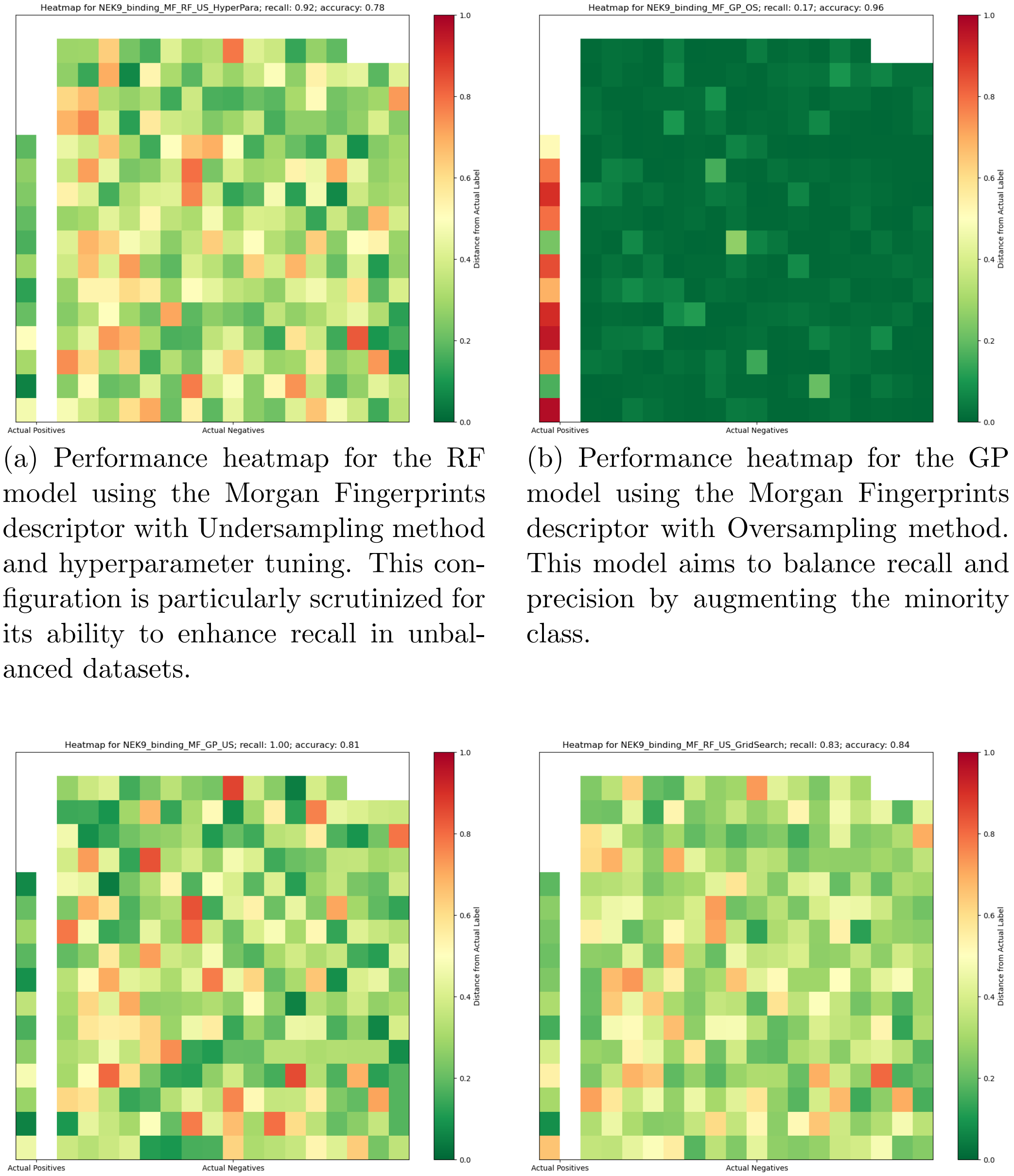


Figure 2: Recall and accuracy scores by model and descriptor for the inhibition dataset.

4.2 Detailed Heatmaps for Binding Dataset

Figures 3a, 3b, 3c, and 3d present heatmaps that show the performance of specific model and method types for the binding dataset. This subsection presents detailed heatmaps depicting the performance of different models and methods applied to the binding dataset. Each heatmap visualizes the recall and accuracy scores, which are essential metrics for evaluating the efficacy of drug discovery models. Green cells within the heatmaps indicate predictions that are closer to real data, signifying a higher accuracy and recall. Conversely, yellow or red cells represent predictions that diverge significantly from the actual data. The greener cells on the left side particularly denote real positives effectively identified by the model, emphasizing the models' capability in capturing true active compounds—a critical aspect for the success of our drug discovery efforts.

Figure 3: Heatmaps of Recall and Accuracy for Different Model Configurations on the Binding Dataset



(a) Performance heatmap for the RF model using the Morgan Fingerprints descriptor with Undersampling method and hyperparameter tuning. This configuration is particularly scrutinized for its ability to enhance recall in unbalanced datasets.

(b) Performance heatmap for the GP model using the Morgan Fingerprints descriptor with Oversampling method. This model aims to balance recall and precision by augmenting the minority class.

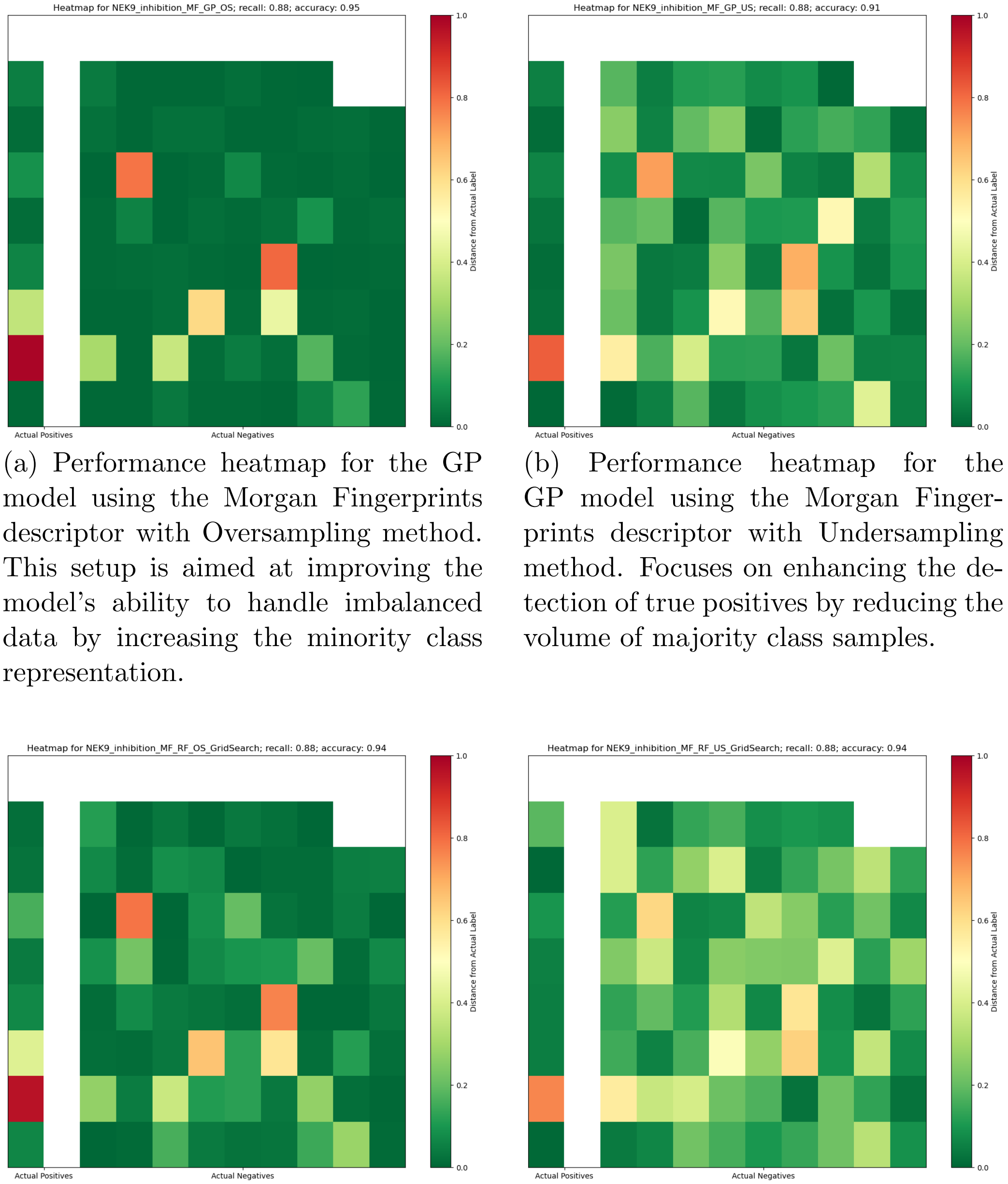
(c) Performance heatmap for the GP model using the Morgan Fingerprints descriptor with Undersampling method. Focuses on reducing false positives while maintaining a strong ability to identify true positives.

(d) Performance heatmap for the RF model using the Morgan Fingerprints descriptor with Undersampling method and grid search for optimal hyperparameter settings. This configuration demonstrates the impact of methodical tuning on model performance in identifying active compounds.

4.3 Detailed Heatmaps for Inhibition Dataset

The following figures (Figure 4a, Figure 4b, Figure 4c, and Figure 4d) display heatmaps for the inhibition dataset, following a similar format to the binding dataset analysis. These heatmaps help visualize which models and methods yield the best performance in terms of recall and accuracy, essential for assessing the effectiveness of drug discovery models. Each heatmap illustrates how different configurations perform under the challenge of identifying inhibitors, a critical task in advancing therapeutic discoveries.

Figure 4: Heatmaps of Recall and Accuracy for Different Model Configurations on the Inhibition Dataset



(a) Performance heatmap for the GP model using the Morgan Fingerprints descriptor with Oversampling method. This setup is aimed at improving the model's ability to handle imbalanced data by increasing the minority class representation.

(b) Performance heatmap for the GP model using the Morgan Fingerprints descriptor with Undersampling method. Focuses on enhancing the detection of true positives by reducing the volume of majority class samples.

(c) Performance heatmap for the RF model using the Morgan Fingerprints descriptor with Oversampling method. This method attempts to balance the dataset to improve both recall and precision effectively.

(d) Performance heatmap for the RF model using the Morgan Fingerprints descriptor with Undersampling method. This configuration is designed to maximize the model's performance in recognizing true active compounds, crucial for effective inhibition prediction.

5 Discussion

This section elaborates on the performance differences observed between various machine learning strategies applied to the drug discovery process for NimA-related kinase (NEK) binders and inhibitors.

5.1 Descriptor Performance

Analysis indicates that Morgan Fingerprints (MF) descriptors outperform Molecular Operating Environment (MOE) descriptors across both binding and inhibition datasets. This superiority in performance is highlighted by higher recall scores achieved by MF descriptors, demonstrating their efficacy in selectively training models to identify active compounds accurately.

5.2 Sampling Methods

Our comparative analysis from Figures 1 and 2 shows that Oversampling (OS) achieves higher accuracy compared to Undersampling (US). This improvement is attributed to OS's ability to mitigate class imbalance by augmenting the minority class, thereby providing a more balanced dataset for model training.

5.3 Dataset-Specific Performance

- **Binding Dataset:** Undersampling (US) is preferred for the binding dataset due to its higher effectiveness in capturing true positives, critical for identifying active binders.
- **Inhibition Dataset:** For the inhibition dataset, Oversampling (OS) is recommended because it reduces uncertainty and maintains high recall, essential for effective inhibitor identification.

5.4 Model Type Considerations

While the Gaussian Process (GP) model generally exhibits superior performance, the use of Random Forest (RF) with Undersampling (US) and the MOE descriptor is beneficial for the binding dataset, indicating that the choice of model and sampling method should be tailored to specific dataset characteristics.

6 Conclusion

The insights gleaned from this research highlight the imperative of adopting targeted machine learning strategies that are specifically tailored to meet the nuanced demands of drug discovery datasets. For the Accelerating Therapeutics for Opportunities in Medicine (ATOM) group, we outline several strategic recommendations based on our findings:

1. **Adopt Morgan Fingerprints (MF) Descriptors:** The superior performance of MF descriptors in terms of recall and accuracy underscores their value over Molecular Operating Environment (MOE) descriptors. We strongly recommend prioritizing MF descriptors in future model training endeavors to enhance the precision of drug candidate identification.

2. **Tailor Sampling Methods to Dataset Needs:** The choice of sampling method should be carefully matched to the specific requirements of the dataset. Oversampling is advisable for datasets where reducing uncertainty is paramount, while Undersampling should be considered when the identification of true positives is more crucial. This strategic approach helps in optimizing model performance and improving prediction accuracy.

3. **Model and Descriptor Synergy:** It is essential to continuously assess and refine the interplay between models and descriptors. This ongoing evaluation is crucial, especially when dealing with complex datasets involved in NEK binder and inhibitor identification, to ensure that the machine learning models are optimally tuned to the characteristics of the data.

Additionally, we have archived the trained models as '.pkl' files within our GitHub repository, making them readily available for the ATOM group's further testing and validation. These models are positioned as a resource to significantly propel the drug discovery initiatives forward.

By integrating these tailored recommendations, the ATOM group is well-placed to boost the efficiency and efficacy of their drug discovery pipelines. Our study equips the team with enhanced capabilities to refine decision-making processes and achieve superior outcomes in their ongoing and forthcoming projects. This proactive approach to machine learning application in drug discovery not only streamlines research efforts but also accelerates the journey from laboratory research to market-ready therapies.

7 Next Steps, Future of The Software Project, or Concluding Remarks

As we look toward the future of this software project, several strategic steps are outlined to enhance its impact and efficacy within the ATOM group and potentially the broader scientific community:

- **Acquire More Training Data:** To improve the model's performance and robustness, we plan to expand the dataset with more diverse chemical compounds. More data will help in refining the models further, especially in reducing overfitting and improving generalization.

- **Exploratory Data Analysis (EDA) on Mistaken Predictions:** Conducting EDA on errors and mistaken predictions can provide insights into where the models may be failing and guide adjustments in feature engineering or modeling approach.

- **Experimentation with Diverse Model Types:** Beyond Gaussian Processes and Random Forests, exploring additional models like deep learning architectures or ensemble methods could uncover more effective approaches for the specific challenges of drug discovery.

- **Maintenance and Further Development:** Continuous development and maintenance will be ensured through regular updates and optimizations based on ongoing testing and feedback from usage within ATOM.

- **Software Availability:** The software and all associated deliverables, including trained models, are hosted on GitHub. This not only facilitates easy access and transparency but also supports collaborative improvements and adaptations by other researchers or developers.

The project's impact extends beyond the immediate improvements in drug discovery processes, promising to accelerate the pace at which new therapeutic agents can be identified and brought to trial, significantly benefiting the medical and pharmaceutical fields.

8 Capstone Project Self-Reflection

Reflecting on the Capstone project management and implementation, this experience has been profoundly enriching, integrating various aspects of software engineering, project management, and team collaboration.

- **Software Project Management:** Throughout the semester, project milestones and deliverables were meticulously tracked using tools like GitHub and Trello. This not only helped in maintaining a clear roadmap but also in adapting to changes efficiently, ensuring that project goals were met on time.

- **Software Engineering Practices:** Best practices in software development were adhered to, including version control, modular coding, and thorough documentation. These practices ensured high-quality software development and ease of maintenance and scalability.

- **Team Collaboration:** Working within a cross-functional team, feedback from peers and supervisors played a critical role in refining the project's direction and outcomes. This collaborative environment was vital in overcoming challenges and leveraging diverse perspectives for innovative solutions.

- **Technical Skills Application:** The project was an opportunity to apply advanced machine learning techniques learned during the MSSE program, particularly in the areas of model optimization and handling imbalanced data.

- **Communication and Negotiation:** Effective communication skills were crucial in negotiating deliverables, presenting ideas, and receiving feedback during professional meetings. These interactions were pivotal in aligning the project with organizational goals and user needs.

- **Risk Management:** Potential risks such as data insufficiency and model overfitting were identified early in the project lifecycle. Proactive strategies were employed to mitigate these risks, ensuring the project's success.

- **Project Updates and Self-Assessment:** Regular project updates and self-assessments were conducted to monitor progress and make necessary adjustments, which were crucial for the project's continuous alignment with its objectives.

This project not only enhanced my technical and managerial skills but also provided valuable insights into the complexities of real-world applications, particularly in the pharmaceutical and healthcare industries.

A Appendix: Technical Documentation and Repository Links

A.1 Repository Overview

This appendix provides detailed information about the Python scripts, Jupyter Notebooks, and datasets utilized in our study on predicting binding and inhibitor compounds for NimA-related kinases (NEK9) using the AMPL software. The repository houses comprehensive scripts for data preprocessing, model training, uncertainty quantification, and result visualization, along with Jupyter Notebooks that demonstrate the step-by-step execution of the analysis.

A.2 Software and Scripts

The project is organized into several modules, each containing Python scripts and Jupyter Notebooks that cover different aspects of the machine learning pipeline:

- **Data Preprocessing Module:** Includes scripts and notebooks for data cleaning, normalization, and feature extraction using both MOE and Morgan Fingerprints descriptors based on SMILES strings.

- **Model Training Module:** Contains scripts and notebooks for configuring, training, and validating Gaussian Process (GP) and Random Forest (RF) models. This module also includes techniques for addressing imbalanced datasets, such as SMOTE, Random Undersampling, and ADASYN.

- **Uncertainty Quantification Module:** Provides scripts and notebooks for calculating uncertainty in model predictions, focusing on confidence intervals from GP models and variance estimates from RF models.

- **Visualization and Analysis Module:** Features scripts and notebooks for generating comprehensive plots and other visual aids to interpret the models' outputs, such as confusion matrices and variance distributions.

A.3 Dataset Documentation

The dataset comprises compound data for NEK9, annotated with binary labels indicating their activity status (active or not active). Data was sourced from publicly available chemical libraries and was processed using the AMPL toolkit for featurization, ensuring relevance and accuracy in model training.

A.4 Repository Link

The complete repository, containing all the Python scripts, Jupyter Notebooks, saved models in '.pkl' format, dataset information, and detailed documentation, is available for public access at:

`https://github.com/TsukiTiger/NEK9_Final.git`

A.5 Reporting to ATOM

Throughout the project, findings were systematically reported to the ATOM consortium during scheduled meetings. These reports included comprehensive updates on model training progress, handling of dataset imbalances, and initial results from uncertainty quantification efforts. Key presentations and meeting notes have been documented and are available in the repository for further review and reference.

A.6 Usage and Replication

To facilitate replication and further exploration, a detailed step-by-step guide is provided in the repository's README file. This guide covers the setup of the computational environment, execution of scripts and notebooks, and detailed interpretation of outputs, enabling researchers to validate and potentially extend the findings with new models or datasets.

References

[1] O. J. Wouters, M. McKee, and J. Luyten, “Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018,” *JAMA*, vol. 323, no. 9, p. 844, Mar. 2020.

[2] S. M. Paul et al., “How to improve RD productivity: the pharmaceutical industry’s grand challenge,” *Nature Reviews Drug Discovery*, vol. 9, no. 3, pp. 203–214, Feb. 2010.

[3] A. Vogelsang and M. Borg, “Requirements Engineering for Machine Learning: Perspectives from Data Scientists,” *IEEE Xplore*, Sep. 01, 2019.

[4] Y. Li et al., “Deep Bayesian Gaussian processes for uncertainty estimation in electronic health records,” *Scientific Reports*, vol. 11, no. 1, Oct. 2021.

[5] Ya Ju Fan et al., “Evaluating point-prediction uncertainties in neural networks for protein-ligand binding prediction,” *Artificial Intelligence Chemistry*, vol. 1, no. 1, pp. 100004–100004, Jun. 2023.

[6] A. J. Minnich et al., “AMPL: A Data-Driven Modeling Pipeline for Drug Discovery,” *Journal of Chemical Information and Modeling*, vol. 60, no. 4, pp. 1955–1968, Apr. 2020.

[7] Carl Edward Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Cambridge, Mass.: MIT Press, 2008.

[8] J. Görtler, R. Kehlbeck, and O. Deussen, “A Visual Exploration of Gaussian Processes,” *Distill*, vol. 4, no. 4, p. e17, Apr. 2019.

[9] C. Do, “The Multivariate Gaussian Distribution,” 2008.

[10] Malte Kuß, *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*, 2006.

[11] R. Golden, *Statistical Machine Learning*. CRC Press, 2020.

[12] Dimitrios Milios, Raffaello Camoriano, Pietro Michiardi, L. Rosasco, and M. Filippone, “Dirichlet-based Gaussian Processes for Large-scale

Calibrated Classification," *neural information processing systems*, vol. 31, pp. 6005–6015, May 2018.

[13] "1.11. Ensemble methods — scikit-learn 0.22.1 documentation," Scikit-learn.org, 2012.

[14] C. H. Yu, S. DiGangi, A. Jannasch-Pennell, and C. Kaprolet, "A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year," *Journal of Data Science*, vol. 8, no. 2, pp. 307–325, Jul. 2021.

[15] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R Journal*, vol. 2, no. 3, 2002.

[16] "GPyTorch's documentation — GPyTorch 0.1.dev97+gf73fa7d documentation," docs.gpytorch.ai.

[17] Scikit-learn, "sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation," Scikit-learn.org, 2018.

[18] "Over-sampling methods — Version 0.10.0," imbalanced-learn.org.

[19] "RandomUnderSampler — Version 0.9.0," imbalanced-learn.org.

[20] A. Lekhtman, "Should I Look at Precision Recall OR Specificity Sensitivity?," Medium, Jan. 10, 2021.

[21] Darko Medin, "Data Science for Drug Discovery Research - Morgan Fingerprints Using Alanine and Testosterone," *Medium*, Nov. 2, 2022.