# Novel multi-objective optimisation methods for multi-label classification in drug discovery

*(Arkia Sabeti's PhD Proposal)*

*Academic email :* as2702@kent.ac.uk | *Personal email:* sabetiarkia97@gmail.com

***Abstract***

*The rising costs and extended periods needed for traditional drug discovery methods call for innovative ways to find new uses for existing medicines, a concept known as drug repurposing. This proposal recommends the use of multi-label algorithms, essential for predicting several drug effects at once, thereby tackling the complexities of diseases and drug interactions. However, multi-labelling presents challenges like computational complexity and the need to manage conflicting goals. Multi-objective optimisation offers a solution by aiming to improve various performance metrics at the same time, easing the problems of multi-label classification. This approach increases the accuracy of identifying drug candidates, supports drug repurposing, and enhances personalised medicine by customising treatments. In conclusion, integrating multi-label classification algorithms enhanced by multi-objective optimisation could greatly speed up drug discovery, lowering both expenses and development times, and leading to safer, more effective treatments.*

## 1. Introduction

In the field of drug discovery and pharmaceutical progress, the use of machine learning-based methods marks a significant development, boosting the search for new therapeutic uses through drug repurposing [1]. Machine learning techniques, which include feature vector-based and similarity-based methods, play a key role in improving predictions of drug-target interactions [1]. Feature vector-based approaches make use of extensive datasets, featuring chemical descriptors of drugs and target sequences, to create feature vectors analysed with conventional machine learning models [1],[2]. Conversely, similarity-based methods rely on information about similarities between drugs and targets, integrating this data into their predictive models. An important example of this type of approach is PREDICT, which uses logistic regression on features coming from drug-drug similarity and disease-disease similarity, drawing on various ontological and interaction data [1], [3].

This shift towards computational methodologies is propelled by the critical need to navigate the increasingly complex and data-intensive landscape of biomedical research. The last few decades have witnessed significant transformations in drug discovery, driven by the rapid collection and analysis of biological data [4]. Despite the promise these advancements offer, the traditional pathways of drug discovery are still beset by inefficiencies, particularly in terms of the significant time and financial costs involved. The process of developing and approving a new drug, often a journey taking nearly 15 years and costing upwards of 800 million dollars [1], emphasises the challenges in achieving a satisfactory return on research and development investments. In this context, drug repurposing, underpinned by the precision of multi-label classification and the strategic framework of multi-objective optimisation, emerges as a cost-effective strategy. Multi-objective optimisation refers to the simultaneous optimisation of multiple conflicting objectives, facilitating a balanced consideration of various performance metrics in multi-label classification. This approach aids in navigating the trade-offs inherent in predictive modelling, ensuring a more nuanced and effective selection of drug candidates [4].

By considerably narrowing the research scope and using supervised learning techniques, this approach offers a faster progression through the drug discovery pipeline [5]. Supervised learning, in particular, excels in navigating the structured domain of biomedical data, making it a fundamental aspect of drug repurposing initiatives. Its ability to predict outcomes based on a deep understanding of drug actions

and interactions highlights the growing synergy between machine learning and pharmaceutical research, marking the beginning of a new era of efficient and economically sustainable drug discovery processes.

The overarching goal of this research is to develop novel methodologies for multi-objective multi-label classification within the context of drug repurposing.

## 2. Background

The growing amount of biomedical data being gathered has greatly increased the need for multi-label algorithms in the field of drug discovery. This large pool of data reveals many insights ready for exploration, showing that drugs, as individual entities, can affect multiple biological targets or have a variety of effects across different systems. Multi-label algorithms excel in this complex situation, allowing for the analysis of complicated datasets not just to understand the full scope of a drug's impact but also to identify synergistic relationships and potential negative effects. This ability for detailed analysis fits perfectly with the characteristics of multi-label data, which differs from the traditional single-label datasets usually dealt with in supervised learning [6].

In traditional supervised learning frameworks, each training example is associated with a singular label $\lambda$ from a distinct set of labels L [6], simplifying the process of predicting outcomes based on learned data relationships. However, the real-world application domains, particularly evident in drug discovery, frequently present scenarios where training examples are tagged with a subset of labels $Y \subseteq L$ [6], categorising them as multi-label. This distinction underscores the complexity and richness of multi-label data, necessitating algorithms that can adeptly handle multiple associations and the interconnectedness of drug effects and biological responses. Thus, the shift from single-label to multi-label data analysis in drug discovery is not merely a technical adjustment but a fundamental enhancement of how we approach the vast and nuanced datasets characteristic of biomedical research.

While multi-label algorithms are vital in dealing with the complexities of drug discovery, they present several challenges that need to be tackled to maximise their benefits. One such challenge is the exponential growth of potential label groupings, which significantly complicates the prediction process [7]. Especially in drug repurposing, where human health and safety are at stake, finding the right balance between a model's accuracy and its simplicity becomes crucial. The aim is to identify attributes that not only enhance the model's ability to predict accurately but also maintain a small set of attributes, keeping the process efficient without losing the model's clarity. This balance is essential in ensuring that the insights gained from data mining algorithms are understandable and usable for end-users, like researchers and clinicians who are looking to discover new therapeutic uses for existing drugs.

Another issues with multi-label classification is the inherent difficulty in managing the interdependencies and correlations between labels [15]. Drugs often target multiple pathways or conditions that are not independent of each other, making the prediction process more complex due to the intricate relationships that need to be modelled accurately.

One further challenge is the imbalance in label distribution [15], a common scenario in drug discovery datasets where some targets or diseases are much more frequently associated with drugs than others. This imbalance can skew the model's learning process, leading to a bias towards more common labels and potentially overlooking rare but important drug-target interactions.

Moreover, evaluating the performance of multi-label classification models poses its own set of difficulties. Unlike single-label classification, where metrics like accuracy are straightforward to calculate and interpret, multi-label scenarios require more nuanced metrics that can account for the correctness of predictions across multiple labels. Measures such as Hamming loss, subset accuracy, and

the F1 score for each label need to be considered, and no single metric can fully capture the model's performance [16], making it challenging to assess and compare the effectiveness of different models.

After addressing and acknowledging the critical challenges, the conclusion is that simultaneous optimisation of multiple quality criteria is not just desirable but necessary. For example, the development of classification models and the selection of attributes for these models encapsulate the dual aim of enhancing predictive accuracy while ensuring the models or attributes remain comprehensible [8]. This duality underscores the multi-objective nature prevalent in many data mining undertakings, dictating that a model's efficacy be measured by a multidimensional vector representing diverse quality criteria, rather than a solitary metric.

Multi-objective optimisation is a strategy in computational analysis where the goal is to optimise two or more conflicting objectives simultaneously. Unlike traditional optimisation techniques that focus on a single objective, multi-objective optimisation acknowledges that in real-world scenarios, objectives often conflict—improving one may worsen another [8].

The combination of multi-objective optimisation with multi-label classification, especially in the context of drug discovery, presents a promising yet relatively underexplored avenue in literature. This gap signifies an opportunity for pioneering research that can significantly contribute to advancing drug repurposing efforts. By melding these two computational strategies, there is potential not only to navigate the challenges presented by multi-label data more effectively but also to pioneer a novel approach that could redefine predictive modelling in drug discovery. As such, this research direction, exploring the synergy between multi-objective optimisation and multi-label classification, offers a fertile ground for originality, promising to advance our understanding and capabilities in the pharmaceutical research landscape.

## 3.Methodology

The methodology at the heart of this research proposal utilises the Pareto approach to tackle multi-objective optimisation, distinct from the traditional practice of transforming a problem with multiple objectives into a single-objective issue for simplification, in the context of multi-label classification.

Originally conceptualised by Francis Ysidro and later extended by Vilfredo Pareto, a solution is considered part of the Pareto set if improving one objective does not deteriorate any other. In the context of multi-objective minimisation, which is pivotal in optimising drug discovery processes, Pareto dominance is defined through the comparison of decision vectors [9]. A decision vector $\upsilon^{\rightarrow}=[\upsilon_1, \upsilon_2, \upsilon_3, \ldots, \upsilon_n]^T$ said to Pareto-dominate another vector $\nu^{\rightarrow} = [\nu_1, \nu_2, \nu_3, \ldots, \nu_n]^T$ if, for all objectives i, the outcome $fi(\upsilon^{\rightarrow})$ is less than or equal to the outcome $fi(\nu^{\rightarrow})$, and there exists at least one objective j where $fj(\upsilon^{\rightarrow})$ is strictly less than $fj(\nu^{\rightarrow})$ [9]. In other words, Pareto dominance establishes a framework where a solution s1 is considered to dominate another solution s2 if s1 is strictly superior in at least one objective without being inferior in any other objective [8]. This concept prioritises the independent evaluation of each objective, maintaining the integrity of multi-objective analysis without reducing it to a single aggregated metric. This criterion ensures that a solution is only considered superior if it offers an unequivocal improvement in at least one objective without any compromise in other objectives.

The reason this proposal suggests using this methodology, is due to its comprehensive ability to address the multifaceted challenges in multi-objective optimisation for multi-label classification.

One of the critical issues with alternatives such as the Minimum Description Length (MDL) principle is the artificial commensurability it introduces between disparate model quality criteria, such as accuracy and size [8]. This commensurability comes at the cost of the complex problem of encoding hypotheses and exceptions into bits, a task that becomes exceedingly difficult as the hypothesis space

expands. The effectiveness of any given encoding is highly application-dependent, making the choice of an appropriate encoding scheme a challenging and often subjective process [10], [8]. The Pareto approach sidesteps these issues by treating model-quality criteria as separate entities, thus respecting their natural non-commensurability and avoiding the pitfalls associated with encoding schemes.

Additionally, the method of using a single-objective optimisation algorithm repeatedly with different weight combinations is both makeshift and not very efficient [11], [12]. This method doesn't take into account the solutions found in earlier attempts, possibly going over solutions that have already been considered and not having a way to ensure a varied set of solutions across the Pareto front. It's especially weak at finding top solutions in the non-smooth areas of the Pareto front, a challenge that the Pareto method naturally overcomes [8].

The concern about how hard it can be to pick the "best" solution from a group of top contenders is lessened by the interactive nature of the discovery process. Having users involved is key, as it makes the subjective nature of selecting the "best" solution less of an issue [13], [14], [8]. Unlike the approach that relies on pre-set weights, which demands early decisions about the importance of different factors without seeing the outcomes, the Pareto method lets users decide after seeing a variety of top solutions. This shows the trade-offs between different criteria, helping users to pick a solution that matches their preferences and knowledge [8].

Combining the Pareto approach's broad, efficient, and user-focused method with the unexplored territory of integrating multi-objective optimisation and multi-label classification in drug discovery, this research takes a novel and promising direction in drug repurposing. The Pareto methodology, despite its complexities, offers a robust framework for the challenges of multi-objective problems. This fusion in drug discovery marks a meaningful step forward, potentially enhancing predictive modelling in the field.

To concretise the methodology within the multi-objective optimisation framework, the initiative plans to integrate the Pareto approach with a variety of classification algorithms. This includes Random Forests, K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and the exploration of additional algorithms as needed. This strategic integration is aimed at leveraging the strengths of these diverse computational techniques to effectively address the research's multifaceted objectives.

The model's performance will be assessed using cross-validation, a well-established method in machine learning for evaluating predictive models by partitioning the data into complementary subsets, training the model on one subset, and validating it on the other. Key multi-label classification measures, serving as the different objectives to be optimised, will include but are not limited to, accuracy, Hamming loss, and the F1 score. These metrics will help in evaluating the efficacy and robustness of the developed multi-objective classification algorithm across various scenarios.

## 4.Data Set and Data preparation

Initial stages of the research will utilise data primarily sourced from the SIDER [17] database for drug side effects and indications, alongside the STITCH [18] database for drug-protein interaction information. The SIDER database, offering a comprehensive repository of marketed medicines and their recorded adverse drug reactions, will provide the class labels (indications) for the multi-label classification problem. Concurrently, the STITCH database will be employed for its detailed drug-protein interaction data, serving as predictive features within the models.

The selection of these databases is based on their proven applicability in related research domains, particularly in drug discovery and repurposing efforts, ensuring relevance and utility in addressing the project’s objectives. Moreover, these datasets are freely available which is not always the case for data in Pharmaceutical context. Data preparation will encompass the extraction of pertinent features and labels from these databases, followed by standard pre-processing procedures, including the

management of missing values, normalisation of numerical data, and encoding of categorical variables. This strategy is aimed at preparing a dataset that is optimally configured for the development and testing of the proposed multi-objective classification algorithm, thus laying a robust foundation for tackling the intricacies of multi-label classification within the drug discovery context.

As the research progresses, the exploration of additional data sources may be undertaken to augment and refine the study. However, for the scope of this proposal, the SIDER and STITCH databases are identified as the primary data sources for the development and evaluation of the proposed methodologies.

The data preparation process will involve standard pre-processing steps to ensure the quality and consistency of the dataset used for model training and testing. This includes handling missing values, normalising data, and encoding categorical variables. The focus will be on preparing a dataset that accurately reflects the complexity and nuances of drug discovery data, laying a solid foundation for the development and evaluation of the proposed classification algorithm.

## 5.Timeline Estimation

Year 1: Data Preparation and Initial Model Development

### Months 1-3: Data Acquisition and Pre-processing

- Source and pre-process data from the SIDER and STITCH databases, including handling missing values, normalising, and encoding categorical variables to prepare for model development.
- Start a literature review on multi-objective optimisation and multi-label classification, focusing on applications of the Pareto approach.

### Months 4-6: Research and Preliminary Model Development

- Continue in-depth literature review on the Pareto approach and multi-label classification.
- Begin developing initial models using basic algorithms like Random Forests, KNN, and SVM.

### Months 7-9: Initial Model Testing and Refinement

- Perform initial tests on developed models to assess performance.
- Refine models based on test results, focusing on improving accuracy and handling of multi-label data.

### Months 10-12: Advanced Model Development

- Integrate advanced algorithms and optimisation techniques into model development.
- Conduct further testing and refinement of models using cross-validation techniques.

Year 2: Model Evaluation and Refinement

### Months 13-15: Comprehensive Model Testing

- Evaluate models using a broader set of metrics such as accuracy, Hamming loss, and the F1 score. Begin the iterative process of refining models based on these comprehensive evaluations.

### Months 16-18: Integration of Pareto Approach

- Further integrate the Pareto approach with multi-label classification strategies. Refine and optimise models to enhance their capability to manage multi-objective optimisation effectively.

### Months 19-21: Evaluation of Model Robustness

- Assess the robustness and scalability of the models across different datasets and scenarios. Adjust models as necessary to address any discovered limitations.

### Months 22-24: Final Model Optimisation

- Apply final optimisations to the models. Prepare for in-depth performance evaluation to ensure the models meet the research objectives.

Year 3: Documentation, Evaluation, and Conclusion

### Months 25-27: In-depth Performance Evaluation

- Conduct a final, thorough evaluation of the models, focusing on their effectiveness in drug discovery and repurposing contexts. Begin drafting findings and methodology sections for the thesis.

### Months 28-30: Thesis Writing and Preparation

- Continue writing the thesis, focusing on discussing the results, implications, and contributions of the research to the field of drug discovery.

### Months 31-33: Thesis Revision and Submission Preparation

- Revise the thesis based on feedback from advisors and peers. Prepare for the submission process, ensuring all components of the thesis are complete and adhere to university guidelines.

### Months 34-36: Defence Preparation and Conclusion

- Finalise thesis and prepare for the defence. Summarise the research contributions, methodologies, and findings in preparation for presentation and defence.

## References


[1]: Manicavasaga, R., Lamichhane, P.B., Kandel, P. and Talbert, D.A. (2022) 'Drug repurposing for rare orphan diseases using machine learning techniques', The International FLAIRS Conference Proceedings, 35.

[2]: Ding, H., Takigawa, I., Mamitsuka, H. and Zhu, S., 2014. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Briefings in bioinformatics, 15(5), pp.734-747.

[3]: Gottlieb, A., Stein, G.Y., Ruppin, E. and Sharan, R., 2011. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Molecular systems biology, 7(1), p.496.

[4]: J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer et al., "Applications of machine learning in drug discovery and development", *Nature reviews Drug discovery*, vol. 18, no. 6, pp. 463-477, 2019.

[5]: Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3(8):673–83.

[6]: Tsoumakas, G., Katakis, I. and Vlahavas, I., 2010. Mining multi-label data. Data mining and knowledge discovery handbook, pp.667-685.

[7]: Aurangzeb, K., Ayub, N. and Alhussein, M., 2021. Aspect based multi-labeling using SVM based ensembler. IEEE Access, 9, pp.26026-26040.

[8]: Freitas, A.A., 2004. A critical review of multi-objective optimization in data mining: a position paper. ACM SIGKDD Explorations Newsletter, 6(2), pp.77-86.

[9]: Ngatchou, P., Zarei, A. and El-Sharkawi, A., 2005, November. Pareto multi objective optimization. In Proceedings of the 13th international conference on, intelligent systems application to power systems (pp. 84-91). IEEE.

[10]: Ross Quinlan, J. and Rivest, R.L., 1989. Inferring decision trees using the minimum description lenght principle. Information and computation, 80(3), pp.227-248.

[11]: D. Corne, K. Deb and P.J. Fleming. The good of the many outweighs the good of the one: evolutionary multi-objective optimization. IEEE Connections Newsletter 1(1), 9-13. IEEE Neural Networks Society, Feb. 2003.

[12]: Deb, K., 2001. Multi-objective optimization using evolutionary algorithms (Vol. 16). John Wiley & Sons.

[13]: Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. eds., 1996, February. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence.

[14]: Brachman, R.J. and Anand, T., 1996. The process of knowledge discovery in databases. In Advances in knowledge discovery and data mining (pp. 37-57).

[15]: Pant, P., Sai Sabitha, A., Choudhury, T. and Dhingra, P., 2019. Multi-label classification trending challenges and approaches. Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, pp.433-444.

[16]: Rainio, O., Teuho, J. and Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14(1), p.6086.

[17]: Kuhn, M., Letunic, I., Jensen, L.J. and Bork, P., 2016. The SIDER database of drugs and side effects. Nucleic acids research, 44(D1), pp.D1075-D1079.

[18]: Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. and Bork, P., 2007. STITCH: interaction networks of chemicals and proteins. Nucleic acids research, 36(suppl_1), pp.D684-D688.