Enhancing ADMET Property Models Performance through Combinatorial Fusion Analysis.

Nan Jiang,[†,§] Mohammed Quazi,[*,‡,§] Christina Schweikert,[¶] D. Frank Hsu,[*,†] Tudor I. Oprea,[*,‡] and Suman Sirimulla[*,‡]

†*Laboratory of Informatics and Data Mining, Department of Computer and Information Science, Fordham University, New York, NY 10023, USA*

‡*Expert Systems, Inc 12730 High Bluff Drive, Suite 100, San Diego, CA 92130, USA*

¶*Division of Computer Science, Mathematics and Science, St. John's University, 8000 Utopia Parkway, Queens, NY 11439, USA*

§ *indicates authors contributed equally*

E-mail: toprea@expertsystems.edu; hsu@fordham.edu; toprea@expertsystems.edu; ssirimulla@expertsystems.edu

Abstract

Accurate prediction of Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties is crucial for drug discovery and development. However, existing computational models for ADMET predictions often lack generalizability and robustness. In this paper, we deployed a Combinatorial Fusion Analysis (CFA) to enhance the performance of ADMET models. Utilizing ADMET benchmark datasets on Therapeutics Data Commons (TDC), we conduct a comprehensive evaluation against traditional and state-of-the-art models. CFA models show superior performance compared

to most of the individual models. The CFA model architecture and the performance of CFA models on TDC and other internal datasets are discussed. This significant enhancement suggests that CFA is a viable tool for improving ADMET model performance, promising faster and more cost-effective drug development pipelines. The code and models trained are available on GitHub at https://github.com/F-LIDM/CFA4DD.

1 Introduction

The field of drug discovery has undergone a transformative evolution in recent years, fueled by advancements in computational methodologies and data-driven approaches. In silico methods have become indispensable tools for accelerating the drug development process by predicting various molecular properties, thereby reducing costs, time, and the need for resource-intensive experimentation. Within this landscape, the application of machine learning and predictive modeling techniques has gained significant attention due to their capacity to accurately forecast crucial drug-related properties [1–3].

One of the key challenges in drug discovery is the accurate prediction of diverse molecular endpoints, such as physico-chemical properties, solubility, permeability, binding affinities and toxicity. The complexity of these endpoints arises from the intricate interplay of multiple factors, necessitating the integration of heterogeneous data sources and diverse molecular descriptors. Part of this complexity relates to the necessity of finding an optimum for the three pillars of drug discovery, diseases, targets and compounds [4]. Within each of these search spaces, an optimal solution needs to be found. Traditional prediction models, while successful to a certain extent, often struggle to capture the inherent nuances of these interdependent relationships. In response, researchers have turned to innovative approaches that leverage the power of ensemble methods and feature fusion techniques.

Combinatorial fusion algorithms represent a promising avenue in addressing the limitations of single-model prediction systems. By amalgamating the strengths of multiple models or descriptors, these algorithms have demonstrated the ability to enhance predictive accu-

racy and robustness. Combinatorial fusion goes beyond the scope of simple model averaging, exploring synergistic interactions among constituent models or descriptors, and exploiting their complementary aspects. This approach provides a means to reduce the risk of overfitting, mitigate bias, and navigate the complex landscape of drug-molecule interactions more effectively.

This paper examines the application of combinatorial fusion algorithms to drug discovery-related endpoint prediction models.

Li and Huang [5] constructed and tested nine unique models: one random forest stacker, two random forests, and six custom gradient boosting models composed of various submodels including random forest and SPGNN: a graph neural network implemented in Hu et al. (2020) [6]. Huang et al. (2020) [7] proposed Deep Purpose which supports training of customized drug-target interaction (DTI) models by facilitating eight compound and seven protein encoders and 50 neural architecture models.

Various graphical models, in particular those based on neural networks, have been used in drug discovery. Kipf and Welling (2017) [8] presented a semi-supervised classification model with graph convolutional neural networks via a localized first-order approximation of spectral graph convolutions. Z. Xiong et al. (2020) [9] proposed a graph neural network architecture for molecular representation. Hu et al. (2020) [6] developed a strategy and self-supervised methods for pre-training Graph Neural Networks (GNN) not only at the level of individual nodes but also as entire graph. Méndez-Lucio et al. [10] presented a molecular foundation model, called MolE, that uses the DeBERTa architecture on molecular graphs and a pretraining strategy.

Virtual screening (VS) of molecular compound libraries using consensus scoring (CS) (or data fusion) has been a viable method for drug discovery and development [11–14]. Charifson et al. (1999) [11] reported that fusion among different scoring methods in VS can perform better than the average of the individual scoring systems. Salim, Holliday, and Willett (2003) [15] reported search results by combining fingerprint-based similarly scoring

systems using data fusion (sum of ranks). Although it was shown that search result can be improved by rank combination with little extra computational cost, there is no clear winner as to which combination gives a consistently higher performance for all search types (e.g. dataset tasks). During the same period, similar phenomena was reported in information retrieval domain [16–18]. Yang et al. (2005) [14] demonstrated, using data from five scoring systems on four protein targets, that combining multiple scoring functions (consensus scoring) improves the enrichment of true positive in virtual screening only if (a) each of the individual scoring functions has relatively high performance and (b) the individual scoring functions are distinctive.

Hsu, Chung, and Kristal (2006) [19] presented Combinatorial Fusion Analysis (CFA) for analyzing and combining multiple scoring systems. CFA characterizes each scoring system A as including a score function s_A (in the Euclidean space), a rank function r_A derived from the score function (in a rank space), and a rank-score function f_A [19, 20]. Both score combination and rank combination are considered w.r.t. their combinatorial complexity and computational efficiency. Information obtained from the scoring characteristics of each scoring system is used to perform system selection and to decide method of combination. In particular, the rank-score functions' graphs, f_A and f_B for scoring systems A and B respectively, has been used to measure the dissimilarity or diversity between scoring systems A and B [19–21].

In this paper, we deploy the CFA algorithm to combine multiple base models which were pretrained using the 22 data sets provided by Therapeutics Data Commons [22]. Section 2 covers methods used in this paper. These include: Section 2.1: Combinatorial Fusion Analysis, Section 2.2: TDC benchmark datasets, and Section 2.3: Training of base and CFA-optimized ADMET-models. Section 3 consists of results and discussion. Result from using CFA with fingerprint descriptors: Morgan and MCFP as well as RDKit 2D descriptors are included in Section 3.1. Section 3.2 includes a summary discussion and suggestion for future work. Section 4 concludes the paper with some remarks. Background regarding the CFA

architecture is included in the Supplemental Information File #1 as: (1) CFA characteristics, (2) Groups, graphs, and Kemeny rank space, and (3) Figure skating judgement: rank v.s. score combination.

2 Methods

2.1 Combinatorial Fusion Analysis (CFA)

Combinatorial Fusion Analysis (CFA) provides ML/AI methods and practices for analyzing combination and fusion of multiple scoring systems (e.g., attributes, algorithms, and models). (See Supplemental Information File #1 for more detailed background information on CFA).

2.1.1 Score Function, Rank Function, and Rank-Score Function

CFA considers each scoring system A as including a score function s_A, a rank function r_A derived from the score function by sorting the score values in descending order, and a rank-score function f_A which depicts the relationship of A in the Euclidean space (s_A) and in the rank space (r_A) [19, 20, 23]. Let $D = \{d_1, d_2, \ldots, d_n\}$ be a set of n objects (subjects, samples, or molecules). Let $N = \{1, 2, \ldots, n\}$ be the set of all positive integers less than or equal to the set of all positive integers n and $\mathbb{R}$ be the set of real numbers. For a scoring system A, the score function $s_A : D \to \mathbb{R}$ is a function from D to $\mathbb{R}$, where the object d_i in D is assigned a real number $s_A(d_i)$ in $\mathbb{R}$. Rank function $r_A : D \to N$ maps each element d_i in D to a natural number $r_A(d_i)$ in N. Suppose $s_A(d_i)$ and $r_A(d_i)$ are the score and rank function on the set D of objects respectively. The rank-score function f_A is defined as: $f_A : N \to \mathbb{R}$ such that

$$f_A(i) = s_A(r_A^{-1}(i)) = (s_A \circ r_A^{-1})(i),\ i \in N \tag{1}$$

which is equivalent to:

$$s_A(d_i) = (f_A \circ r_A)(d_i) = f_A(r_A(d_i)),\ d_i \in D \tag{2}$$

Rank-score function f_A of the scoring system A was defined by Hsu, Shapiro, and Taksa [18, 24] in the study of information retrieval systems. Due to its generality and independence of application domains, it has since been used to measure the dissimilarity/diversity between scoring systems in the duality of Euclidean space and rank space [18, 20, 21, 23, 25]. In particular, cognitive diversity (or rank-score diversity) was defined and widely used in a variety of domains [18, 20, 21, 23].

2.1.2 Score and Rank Combination

Given two scoring systems A and B, their score functions s_A and s_B, and their derived rank functions r_A and r_B, we are able to obtain their rank-score functions f_A and f_B respectively. Score values of A and B are normalized linearly to the interval $[0, 1]$ respectively. Cognitive diversity (a.k.a: rank-score diversity) between scoring systems A and B, $cd(A, B)$, proposed by Hsu et al. [18–21, 24] measures the dissimilarity (or diversity) of scoring systems A and B:

$$cd(A, B) = d(f_A, f_B) = \left(\frac{\sum_{i=1}^{n} (f_A(i) - f_B(i))^2}{n} \right)^{\frac{1}{2}} \tag{3}$$

where n is the number of data items in D.

Suppose we have t scoring systems $A_1, A_2, \ldots, A_t$ on the data set D of n items. The diversity strength of the scoring system A_j, $ds(A_j)$ is defined to be [23, 26, 27]:

$$ds(A_j) = \frac{\sum_{i=1, i \neq j}^{t} cd(A_j, A_i)}{t - 1} \tag{4}$$

On the other hand, the performance strength is simply taken as the performance of the scoring system A_j, $p(A_j)$, which is measured as the correlation, accuracy, precision, or AUROC depending on the data set and the intended task.

For the combination of the t scoring systems $A_1, A_2, \ldots, A_t$, we consider three types of combination: average combination (AC), weighted combination by diversity strength (WCDS), and weighted combination by performance (WCP). Since every scoring system

has a score function and a rank function, we have the following score functions for AC, WCDS, and WCP:

For (a) average combination (AC) of score combination (SC) or rank combination (RC), the score function of the score combination s_{SC} and of the rank combination s_{RC} are:

$$s_{SC}(d_i) = \frac{\sum_{j=1}^{t} s_{A_j}(d_i)}{t} \text{ and } s_{RC}(d_i) = \frac{\sum_{j=1}^{t} r_{A_j}(d_i)}{t} \tag{5}$$

For (b) weighted combination by diversity strength (WCDS), the score function of the score combination and of the rank combination are:

$$s_{SC}(d_i) = \frac{\sum_{j=1}^{t} ds(A_j)(s_{A_j}(d_i))}{\sum_{j=1}^{t} ds(A_j)} \text{ and } s_{RC}(d_i) = \frac{\sum_{j=1}^{t} (1/ds(A_j))(r_{A_j}(d_i))}{\sum_{j=1}^{t} (1/ds(A_j))} \tag{6}$$

For (c) weighted combination by performance (WCP), the score function of the score combination and of the rank combination are:

$$s_{SC}(d_i) = \frac{\sum_{j=1}^{t} p(A_j)(s_{A_j}(d_i))}{\sum_{j=1}^{t} p(A_j)} \text{ and } s_{RC}(d_i) = \frac{\sum_{j=1}^{t} (1/p(A_j))(r_{A_j}(d_i))}{\sum_{j=1}^{t} (1/p(A_j))} \tag{7}$$

when $p(A_j)$ is the higher the better. However, if $p(A_j)$ is the lower the better, the score function of the score combination and of the rank combination are:

$$s_{SC}(d_i) = \frac{\sum_{j=1}^{t} (1/p(A_j))(s_{A_j}(d_i))}{\sum_{j=1}^{t} (1/p(A_j))} \text{ and } s_{RC}(d_i) = \frac{\sum_{j=1}^{t} p(A_j)(r_{A_j}(d_i))}{\sum_{j=1}^{t} p(A_j)} \tag{8}$$

2.2 TDC benchmark datasets

To benchmark our approach, we used the ADMET datasets from Therapeutics Data Commons (TDC). TDC serves as a comprehensive hub for drug discovery research, offering a wide array of curated datasets and benchmarks to facilitate scientific innovation. Therapeutics Data Commons (TDC) also offers a unique leaderboard feature that serves as a real-time assessment platform for various algorithms in drug discovery. Researchers can submit their

models for tasks like ADMET prediction and see how they stack up against existing benchmarks. Metrics tailored to each specific research problem are used to rank the algorithms, offering an instant, comparative view of model performance. In this paper, we used the 22 ADMET benchmark datasets to evaluate our approach and ranked their performance against the leaderboard.

2.3 Training of Base and CFA-optimized ADMET Models

For this study, we made use of the 22 datasets available in the Therapeutics Data Commons (TDC) specifically related to ADMET properties. The ADMET group within TDC provided a reliable and diverse range of data, ensuring robustness in our experimental evaluation.We employed three techniques to generate molecular features for the representation of the compounds in these 22 datasets: 1) Morgan Circular Fingerprints (Configured with a radius of 2 and a bit vector length of 1024), 2) RDKIT 2D Molecular Descriptors, and 3) MCFP, a proprietary fingerprints developed by Hassan, Sirimulla, and Oprea. Five algorithms are used as base models (A, B, C, D, E) = (XGB, RF, SVM, ADB, CNN) where XGB is an implementation of Gradient Boosted decision trees [28], RF is Random Forest [29], SVM is Support Vector Machine with linear kernel [30], ADB is AdaBoost with 300 estimators and learning rate = 1 [31], and CNN is Convolutional Neural Networks with 1 input, 2 hidden, and 1 output layer, Sigmoid activation function for output and ReLu activation function for other layers and Adam optimizer [32]. For model evaluation, we compute the following measurements: Mean Absolute Error (MAE) and Spearman rho in the 9 regression cases. For the 13 classification cases, we use AUROC (area under the ROC curve), and AUPRC (area under precision-recall) as metrics.

In the training and testing process, we follow TDC's protocol of splitting the data set into 20% for testing, and 80% for training and validation (70% v.s. 10%) [22]. The test data is fixed (by TDC format) while the 70%/10% split of the remaining 80% is seeded randomly five times using random split, scaffold split, temporal split, cold-state split, and

combination split. Performance of the model is the average of the performance for the test data. We refrained from extensive hyperparameter tuning using scikit-learn's GridSearch cross-validation, relying instead on scikit-learn's RandomizedGridSearch cross-validation for XGB (parameters = booster, learning rate, maximum depth, minimum child weight, number of estimators, lambda, alpha, sample by tree, and gamma) and SVM (parameters = C and gamma) algorithms, and default settings for the others.

We use three types of combination: average combination (AC), weighted combination by diversity strength (WCDS), and weighted combination by performance (WCP). Five base models for each feature set led to $(2^5 - 1 - 5) \times 3 = 78$ combined models. Each of the 78 combined models is obtained in two different ways (score combination and rank combination).

3 Results and Discussion

3.1 Results

CFA is robust across all five categories of TDC ADMET data sets. It achieves top rank in 4 of the 22 data sets in ADMET category of TDC. CFA has best results in data sets E.1 and E.2, which uses Spearman's rho to evaluate the performance of the modeling result. The evaluation of modeling results with Spearman rho is done on the rank space. Since the five base models are diverse as evidenced by rank-score function graphs (see figures in supplemental information file #2), CFA results using rank combination is better than score combination [18] where there is diversity between these models. CFA is robust w.r.t. small and large data size. It achieves all rankings within the top 6 with data size as small as 475 (T.4: DILI) to data size as big as 13,130 (M.5: CYP2D6_Veith). The six data sets which CFA ranks in the TDC leaderboard over 6 are A.1: Caco2_wang (8/17), A.5: Lipophilicity (14/15), A.6: Solubility (9/14), M.4 CYP2C9_Veith (9/16), M.6: CYP3A4_Veith (8/15), and T.1: LD50_Zhu (7/17). Supplemental Information file #2 consists of 22 tables (S.I. tables) each representing the performance of the 78 score or rank combinations of the com-

bined models and the five single models under each of the five seeding cases. For each of the 22 dataset tasks, five rank-score function graphs (S.I. figures) depicting the model diversity are also included.

3.1.1 CFA with Fingerprints descriptors

Initially, we performed CFA with models developed from Morgan and MCFP fingerprints as descriptors and five base models (A: XGB, B: RF, C: SVM, D: ADB, and E: CNN). The performance of CFA with these combination is shown in Table 1. Table 1 lists performance evaluation metric, the best performance on the leaderboard, CFA performance, and the CFA ranking on the leaderboard for each of the 22 data sets and their tasks. Table 4 lists the performance and diversity strength of each model.

3.1.2 CFA with RDKit 2D descriptors

Next, we explored RDKit 2D descriptors as our features on 11 data sets. The ranking has improved in 6/11 data sets. The leaderboard ranking of CFA-optimized models with RDKit 2D Descriptors is shown in Table 3. The CFA with RDKit 2D Descriptors achieves rankings within the top 6 in 8/11 data sets and falls out of the top 6 in only three data sets (A.3: Pgp_Broccatelli, E.2: Clearance_Hepatocyte_AZ, and T.4: DILI). One of our future works is to continue CFA work using RDKit 2D Descriptors as the encoding scheme.

Table 1: Performance of CFA on ADMET models in the TDC Leaderboard

Category	Dataset	Metric	Leaderboard Top Score*	CFA Score	FP Method	CFA Best Ranking**
Absorption	1.Caco2_wang	MAE↓	0.276	0.335	MCFP	8 / 17
	2.HIA_Hou	AUROC↑	0.989	0.981	MCFP	4 / 16
	3.Pgp_Broccatelli	AUROC↑	0.938	0.928	RDKit 2D	6 / 16
	4.Bioavailability_Ma	AUROC↑	0.748	0.746	MCFP	2 / 16
	5.Lipophilicity_AstraZeneca	MAE↓	0.467	0.626	MCFP	14 / 15
	6.Solubility_AqSolDB	MAE↓	0.761	0.939	MCFP	9 / 14
Distribution	1.BBB_Martins	AUROC↑	0.916	0.920	RDKit 2D	1 / 21
	2.PPBR_AZ	MAE↓	7.526	8.680	MCFP	6 / 15
	3.VDss_Lombarado	Spearman rho↑	0.713	0.628	Morgan	3 / 15
Metabolism	1.CYP2C9_Substrate	AUPRC↑	0.441	0.417	RDKit 2D	5 / 17
	2.CYP2D6_Substrate	AUPRC↑	0.736	0.704	RDKit 2D	4 / 15
	3.CYP3A4_Substrate	AUROC↑	0.662	0.667	MCFP	1 / 16
	4.CYP2C9_Veith	AUPRC↑	0.859	0.751	Morgan	9 / 16
	5.CYP2D6_Veith	AUPRC↑	0.790	0.664	MCFP	6 / 15
	6.CYP3A4_Veith	AUPRC↑	0.916	0.855	Morgan	8 / 15
Excretion	1.Half_life	Spearman rho↑	0.562	0.576	Morgan	1 / 16
	2.Clearance_Hepatocyte	Spearman rho↑	0.498	0.536	Morgan	1 / 14
	3.Clearance_Microsome	Spearman rho↑	0.630	0.625	RDKit 2D	3 / 16
Toxicity	1.LD50_Zhu	MAE↓	0.552	0.630	MCFP	7 / 17
	2.hERG	AUROC↑	0.880	0.875	RDKit 2D	2 / 16
	3.AMES	AUROC↑	0.871	0.852	MCFP	4 / 15
	4.DILI	AUROC↑	0.925	0.919	MCFP	2 / 16

* TDC ADMET benchmark group leaderboard ranking is accessed on November 26, 2023. Same for Table 2 and 3.
** Numbers in red indicate ranking within top 6. Same as in Table 2 and 3.

Table 2: Performance of CFA (Morgan and MCFP) on ADMET models in the TDC Leaderboard

Category	Dataset	Metric	CFA + Morgan	CFA Ranking with Morgan	CFA + MCFP	CFA Ranking with MCFP
Absorption	1.Caco2_wang	MAE↓	0.417	12 / 17	0.335	8 / 17
	2.HIA_Hou	AUROC↑	0.960	10 / 16	0.981	4 / 16
	3.Pgp_Broccatelli	AUROC↑	0.917	8 / 16	0.924	6 / 16
	4.Bioavailability_Ma	AUROC↑	0.685	5 / 16	0.746	2 / 16
	5.Lipophilicity_AstraZeneca	MAE↓	0.654	14 / 15	0.626	14 / 15
	6.Solubility_AqSolDB	MAE↓	1.173	14 / 14	0.939	9 / 14
Distribution	1.BBB_Martins	AUROC↑	0.891	12 / 21	0.907	7 / 21
	2.PPBR_AZ	MAE↓	8.734	6 / 15	8.680	6 / 15
	3.VDSS_Lombarado	Spearman rho↑	0.628	3 / 15	0.561	6 / 15
Metabolism	1.CYP2C9_Substrate	AUPRC↑	0.400	6 / 17	0.358	15 / 17
	2.CYP2D6_Substrate	AUPRC↑	0.684	7 / 15	0.677	7 / 15
	3.CYP3A4_Substrate	AUROC↑	0.642	4 / 16	0.667	1 / 16
	4.CYP2C9_Veith	AUPRC↑	0.751	9 / 16	0.749	9 / 16
	5.CYP2D6_Veith	AUPRC↑	0.660	6 / 15	0.664	6 / 15
	6.CYP3A4_Veith	AUPRC↑	0.855	8 / 15	0.853	8 / 15
Excretion	1.Half_Life_Obach	Spearman rho↑	0.576	1 / 16	0.312	8 / 16
	2.Clearance_Hepatocyte_AZ	Spearman rho↑	0.536	1 / 14	0.411	9 / 14
	3.Clearance_Mircrosome_AZ	Spearman rho↑	0.572	9 / 16	0.566	10 / 16
Toxicity	1.LD50_Zhu	MAE↓	0.632	7 / 17	0.630	7 / 17
	2.hERG	AUROC↑	0.833	7 / 16	0.832	7 / 16
	3.AMES	AUROC↑	0.837	7 / 15	0.852	4 / 15
	4.DILI	AUROC↑	0.863	11 / 16	0.919	2 / 16

Table 3: CFA Performance with RDKit 2D Descriptors

Dataset	Metric	Leaderboard topper	CFA using RDKit 2D	CFA Ranking
A.2 HIA_Hou	AUROC↑	0.989	0.982	4 / 16
A.3 Pgp_Broccatelli	AUROC↑	0.938	0.922	7 / 15
A.4 Bioavailability_Ma	AUROC↑	0.748	0.736	3 / 16
D.1 BBB_Martins	AUROC↑	0.916	0.920	1 / 21
M.1 CYP2C9_Substrate	AUPRC↑	0.441	0.417	5 / 17
M.2 CYP2D6_Substrate	AUPRC↑	0.736	0.704	4 / 15
M.3 CYP3A4_Substrate	AUROC↑	0.662	0.652	2 / 16
E.2 Clearance_Hepatocyte_AZ	Spearman rho↑	0.498	0.42	8 / 14
E.3 Clearance_Microsome_AZ	Spearman rho↑	0.630	0.625	3 / 16
T.2 hERG	AUROC↑	0.880	0.875	2 / 16
T.4 DILI	AUROC↑	0.925	0.886	8 / 16

Table 4: The Performance and Diversity Strength of Best CFA Combined models

Dataset	Average Performance of Individual Model*	Average Diversity Strength of Individual Model	Best CFA Combined Model	Second Best CFA Combined Model	FP Method
A.1	A: 0.405; B: 0.360; C: 0.364; D: 0.529; E: 1.026	A: 1.330; B: 1.094; C: 1.414; D: 2.970; E: 1.038	BC_ps: 0.335	BC: 0.335	MCFP
A.2	A: 0.893; B: 0.929; C: 0.942; D: 0.981; E: 0.887	A: 1.004; B: 0.681; C: 0.666; D: 1.342; E: 1.396	D: 0.981	DE_ps: 0.979	MCFP
A.3	A: 0.800; B: 0.899; C: 0.858; D: 0.903; E: 0.920	A: 0.595; B: 0.727; C: 0.586; D: 0.698; E: 1.529	ABE_ds: 0.928	BE_ds: 0.927	RDKit 2D
A.4	A: 0.718; B: 0.735; C: 0.648; D: 0.705; E: 0.698	A: 1.184; B: 0.654; C: 1.287; D: 1.281; E: 2.538	AB_ps: 0.746	AB: 0.746	MCFP
A.5	A: 0.640; B: 0.711; C: 0.676; D: 0.819; E: 0.739	A: 2.964; B: 3.198; C: 7.158; D: 8.351; E: 8.060	AC_ps: 0.626	AC: 0.626	MCFP
A.6	A: 0.945; B: 1.010; C: 1.097; D: 1.409; E: 1.184	A: 14.912; B: 13.305; C: 13.661; D: 31.257; E: 35.228	ABC_ps: 0.939	A: 0.945	MCFP
D.1	A: 0.868; B: 0.892; C: 0.867; D: 0.833; E: 0.884	A: 0.924; B: 0.534; C: 0.556; D: 0.945; E: 2.001	BCE_ps: 0.920	BCE: 0.919	RDKit 2D
D.2	A: 9.226; B: 9.740; C: 9.021; D: 13.461; E: 14.752	A: 5.266; B: 6.585; C: 14.973; D: 8.711; E: 10.755	AC: 8.680	AC_ps: 8.741	MCFP
D.3	A: 0.348; B: 0.525; C: 0.582; D: 0.321; E: -0.017	A: 5.835; B: 10.073; C: 8.170; D: 7.312; E: 6.027	BC_ds_r: 0.628	BC_r: 0.626	Morgan
M.1	A: 0.317; B: 0.382; C: 0.353; D: 0.313; E: 0.417	A: 1.880; B: 2.244; C: 3.256; D: 4.580; E: 2.703	E: 0.417	BE: 0.415	RDKit 2D
M.2	A: 0.563; B: 0.679; C: 0.568; D: 0.686; E: 0.659	A: 0.634; B: 0.763; C: 0.936; D: 1.417; E: 1.520	BCD: 0.704	BCD_ps: 0.703	RDKit 2D
M.3	A: 0.621; B: 0.662; C: 0.608; D: 0.649; E: 0.563	A: 1.403; B: 0.771; C: 1.143; D: 0.739; E: 2.905	ABD_ps: 0.667	ABD: 0.667	MCFP
M.4	A: 0.739; B: 0.713; C: 0.728; D: 0.571; E: 0.687	A: 36.56; B: 43.81; C: 33.74; D: 108.50; E: 78.73	ABCD: 0.751	ABCD_ps: 0.751	Morgan
M.5	A: 0.649; B: 0.626; C: 0.528; D: 0.538; E: 0.509	A: 30.594; B: 35.394; C: 70.681; D: 84.129; E: 43.781	ABC_ps: 0.664	ABC: 0.664	MCFP
M.6	A: 0.847; B: 0.822; C: 0.841; D: 0.718; E: 0.810	A: 37.01; B: 40.67; C: 34.58; D: 110.74; E: 79.13	ABCD_ps: 0.855	ABCD: 0.855	Morgan
E.1	A: 0.352; B: 0.257; C: 0.474; D: 0.199; E: -0.039	A: 2.365; B: 7.467; C: 2.731; D: 4.032; E: 3.314	ABCD_ds_r: 0.576	ABC_r: 0.574	Morgan
E.2	A: 0.350; B: 0.323; C: 0.396; D: 0.278; E: -0.002	A: 3.190; B: 4.774; C: 3.896; D: 8.294; E: 4.203	ABCDE_r: 0.536	ABCDE_ds_r: 0.536	Morgan
E.3	A: 0.515; B: 0.502; C: 0.609; D: 0.527; E: 0.618	A: 0.607; B: 0.559; C: 0.925; D: 0.627; E: 1.231	CE_r: 0.625	CE_ps_r: 0.625	RDKit 2D
T.1	A: 0.647; B: 0.635; C: 0.684; D: 0.765; E: 0.701	A: 7.798; B: 7.900; C: 11.151; D: 18.333; E: 19.091	ABE_ps: 0.630	ABE: 0.631	MCFP
T.2	A: 0.815; B: 0.803; C: 0.767; D: 0.811; E: 0.850	A: 0.394; B: 0.368; C: 0.485; D: 0.419; E: 1.090	DE: 0.875	ACDE: 0.875	RDKit 2D
T.3	A: 0.835; B: 0.844; C: 0.812; D: 0.759; E: 0.746	A: 10.735; B: 18.572; C: 13.835; D: 23.861; E: 16.525	ABC_ds: 0.852	ABC_ps: 0.851	MCFP
T.4	A: 0.870; B: 0.897; C: 0.851; D: 0.884; E: 0.809	A: 0.660; B: 0.566; C: 0.455; D: 0.794; E: 2.156	ABC_ps: 0.919	ABC: 0.918	MCFP

Note: 1. A: XGB; B: RF; C: SVM; D: ADB; E: CNN.
2. The suffix "ps" means weighted combination by performance strength.
3. The suffix "ds" stands for weighted combination by diversity strength.
4. "r" represents rank combination while all other combinations without "r" indicate score combinations.

Figure 1: RSC Graphs for VDSS_Lombardo

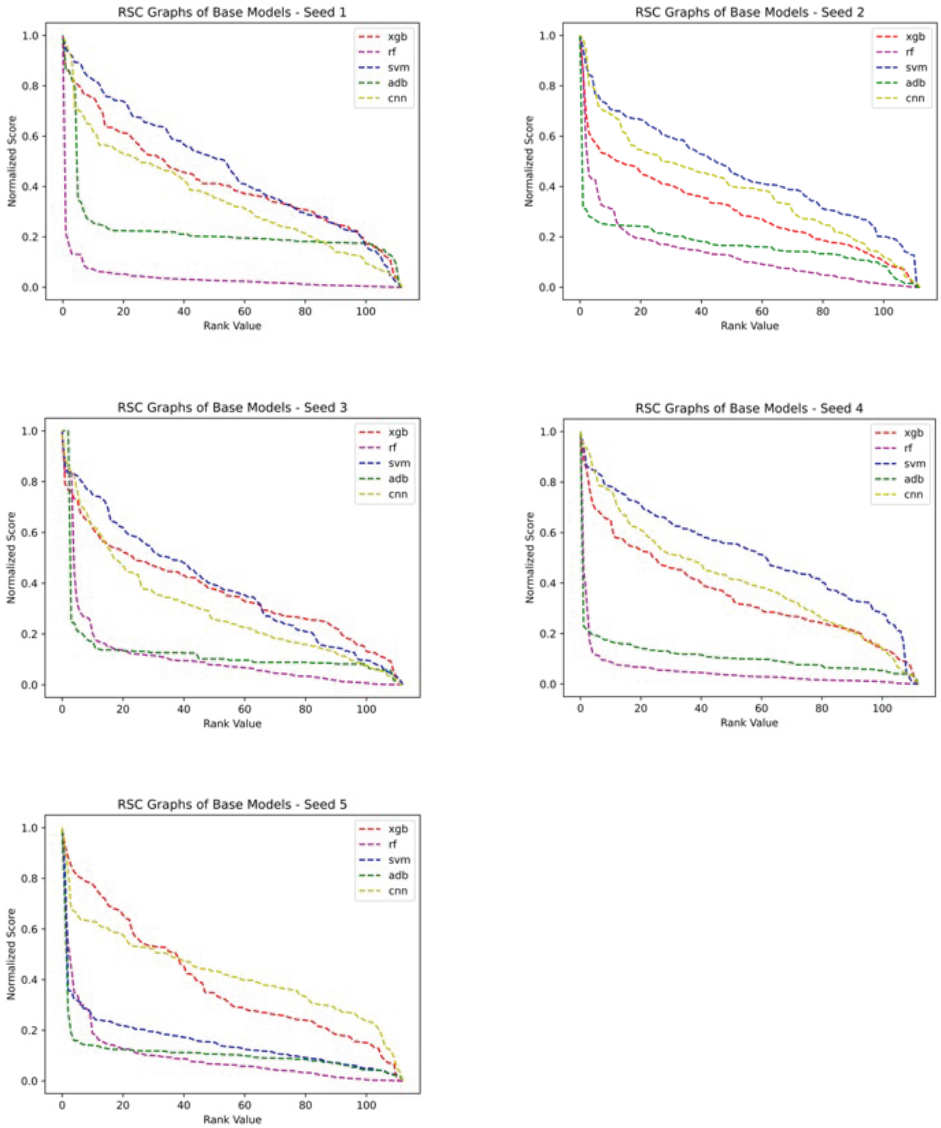

Figure 2: RSC Graphs for Half_Life_Obach

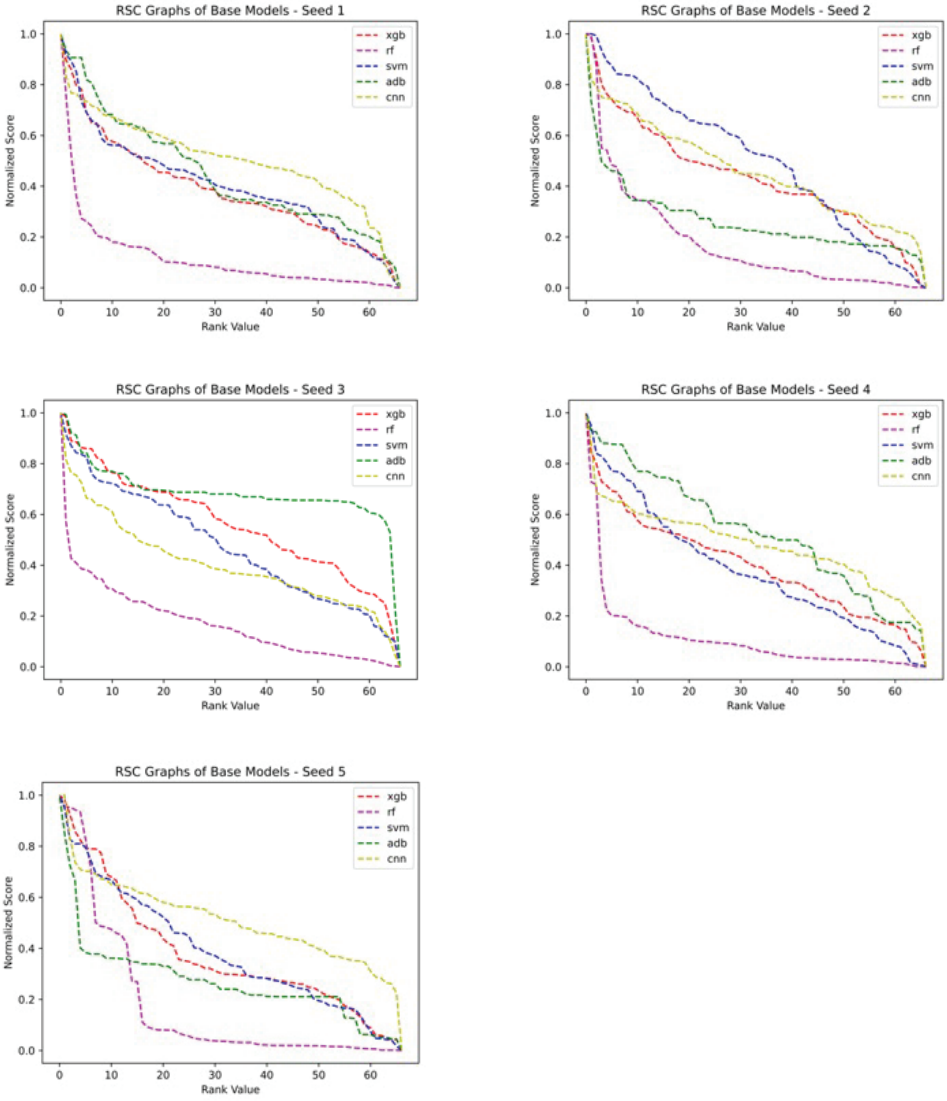

Table 5: CFA + Morgan Performance on VDSS_Lombardo and Half_Life_Obach

Dataset	Metric	# Params	Problem Type	Combination Type	Metric Score
VDSS_Lombardo	Spearman rho↑	1024	Regression	Rank combination	seed_1: 0.649 seed_2: 0.642 seed_3: 0.628 seed_4: 0.630 seed_5: 0.589 average: 0.628
Half_Life_Obach	Spearman rho↑	1024	Regression	Rank combination	seed_1: 0.567 seed_2: 0.543 seed_3: 0.610 seed_4: 0.589 seed_5: 0.571 average: 0.576

3.2 Discussion

Results that improved model prediction of ADMET properties using CFA are given in Tables 1, 3 and 4, as well as the S.I. tables and figures in supplemental information file #2, and summarized in Section 3.1. In this section, we offer some points of discussion together with possible future work.

3.2.1 Rank vs. Score Combination

CFA provides both score combination in the Euclidean space $\mathbb{R}_n$ and rank combination in the Bubble sort Cayley graph space B_n (with no tie rankings) and in the Kemeny rank space (with tie rankings allowed) (see S.I. file #1 Section 2 for Bubble Sort Cayley graph space B_n and Kemeny rank space K_n). However, in Tables 1, 3 and 4, and S.I. tables in the supplemental information file #2 on the 22 data sets and tasks ([22] and TDC, rank combinations are performed on the four data sets (D3, E1, E2, and E3) to match with the metrices used by these four data sets. We note that CFA achieves the #1 spot in two of these respective leaderboards. On the other hand, out of the other 18 score combination results, CFA achieves top 6 spot in 12 data sets.

It was shown ([18, 24]) that rank combination of scoring system A and B can be better than score combination when there exists diversity (or cognitive diversity, [20, 21, 23]) between rank-score functions f_A and f_B. One of our future works is to perform rank combination after converting the score function into rank function in the 18 data sets. We expect the results from the rank combination to be better as there do exist cognitive diversity between these models (see S.I. tables in the supplemental information file #2).

3.2.2 Cognitive Diversity (CD)

The concept of cognitive diversity(or rank-score diversity) between models(scoring systems) A and B is defined as the variation or dissimilarity between two normalized rank-score function of A and B (see Equation (3) in Section 2.1.2). It is an integral core part of CFA.

CD can not only help us decide the method of combination but also help us find a better way to assemble a group of diverse yet relatively good base models. As we can see from tables 4 and 5 for the datasets D.3 (VDSS) and E.1 (Half_Life), the best combinations are from weighted rank combination by diversity strength of RF and SVM and weighted rank combination by diversity strength of XGB, RF, SVM and ADB respectively.

Examining the rank-score function graphs on Figures 1 and 2, We see that models RF and SVM do have large CD between them. Similar results are reflected by the other 20 data sets if rank combination is used.

3.2.3 Learning and Modeling

The CFA framework in this paper considers both statistical and computational approaches of learning and modeling (see Section 1 in supplemental information file #1) according to the nature and content of the data sets and their application domains. In computational learning and modeling, and in particular in molecular science, B. Dou et al. [33] discussed machine learning methods for small data challenges and offered several suggestions for improving the prediction power of ML/DL models w.r.t. small data sets including the combination of ML and DL models. In the CFA framework, all the six combination methods (see Figure 3 in Section 3 of supplemental information file #1) do apply to the five base models (chosen w.r.t. their diversity regardless of them being ML or DL models). For instance, one of the five base models chosen in the first empirical example is CNN, a DL model (Section 2.3). On the other hand, even though the model SVM is less considered in drug discovery, or performance of single SVM may not be high, we include SVM as one of the five base models because it provides good diversity (i.e., cognitive diversity) and hence contributes to enhancing prediction power (S.I. tables in Supplemental Information file #2).

3.2.4 Flexibility of CFA in Combining Models

The salient characteristic of Combinatorial Fusion Analysis (CFA) lies in its inherent flexibility, as it is agnostic of specific features and machine learning algorithms. This allows for the combination of models developed using diverse feature sets and ML algorithms, thereby alleviating researchers from the often laborious task of meticulously selecting an optimal feature set or a singular machine learning algorithm tailored to a specific problem. In this study, we employed three distinct feature sets in conjunction with five different ML algorithms. While the scope of this study focused on the application of CFA at the machine learning algorithms level for each of the three feature sets, it is worth noting that CFA also affords the opportunity to combine models developed using distinct feature sets. This flexibility underscores the versatility of CFA in tailoring model combinations to the specific requirements of the data analysis task at hand.

4 Conclusions

CFA provides efficient algorithms to combine a moderate group of relatively good and diverse models. It takes advantage of combining multiple good and diverse base models using both score and rank combination as well as three types of combinations: AC, WCDS, and WCP. By leveraging ADMET benchmark datasets from Therapeutics Data Commons (TDC), we compared the CFA-optimized models with state-of-art ADMET models. Remarkably, CFA-optimized models clinched the top spot on TDC leaderboard in four of these datasets and maintained a position within the top six ranks in 16 datasets. The integration of CFA into ADMET modeling not only elevates predictive accuracy but also promises to expedite the drug development cycle, thereby enabling more efficient and cost-effective solutions.

Supporting Information Available

The following files are available free of charge.

- Filename: Supplemental Information File #1
- Filename: Supplemental Information File #2

Data and Code availability

The code and protocols used in this paper are available on the GitHub at https://github.com/F-LIDM/CFA4DD. The python package for the CFA is available through pip install and can be installed using the command pip install cfanalysis. More information on python package is avaiable at https://pypi.org/project/cfanalysis/.

References

(1) Klambauer, G.; Hochreiter, S.; Rarey, M. Machine learning in drug discovery. *Journal of Chemical Information and Modeling* **2019**, *59*, 945–946.

(2) Wei, G.-W.; Zhu, F.; Merz, K. M. Editorial on Machine Learning. *Journal of Chemical Information and Modeling* **2022**, *62*, 3941–3941.

(3) Merz, K. M.; Wei, G.-W.; Zhu, F. Machine Learning in Bio-cheminformatics. *Journal of Chemical Information and Modeling* **2023**, *63*, 1–1.

(4) Hasselgren, C.; Oprea, T. I. Artificial Intelligence for Drug Discovery: Are We There Yet? *Annual Review of Pharmacology and Toxicology* **2024**, *64*, null, PMID: 37738505.

(5) Li, A.; Huang, D. OCE-TDC/Report.pdf at main Oloren-AI/OCE-TDC. *GitHub*

(6) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* **2019**,

(7) Huang, K.; Fu, T.; Glass, L. M.; Zitnik, M.; Xiao, C.; Sun, J. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* **2020**, *36*, 5545–5547.

(8) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**,

(9) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H., et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry* **2019**, *63*, 8749–8760.

(10) Méndez-Lucio, O.; Nicolaou, C.; Earnshaw, B. MolE: a molecular foundation model for drug discovery. *arXiv preprint arXiv:2211.02657* **2022**,

(11) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry* **1999**, *42*, 5100–5109.

(12) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047–1055.

(13) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

(14) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *Journal of Chemical Information and Modeling* **2005**, *45*, 1134–1146.

(15) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *Journal of chemical information and computer sciences* **2003**, *43*, 435–442.

(16) Ng, K. B.; Kantor, P. B. Predicting the effectiveness of naive data fusion on the basis of system characteristics. *Journal of The American Society for Information Science* **2000**, *51*, 1177–1189.

(17) Belkin, N. J.; Kantor, P.; Fox, E. A.; Shaw, J. A. Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management* **1995**, *31*, 431–448.

(18) Hsu, D. F.; Taksa, I. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval* **2005**, *8*, 449–480.

(19) Hsu, D. F.; Chung, Y.-S.; Kristal, B. S. Combinatorial fusion analysis: methods and practices of combining multiple scoring systems. *Advanced Data Mining Technologies in Bioinformatics* **2006**, 32–62.

(20) Hsu, D. F.; Kristal, B. S.; Schweikert, C. Rank-score characteristics (RSC) function and cognitive diversity. Brain Informatics: International Conference, BI 2010, Toronto, ON, Canada, August 28-30, 2010. Proceedings. 2010; pp 42–54.

(21) Hsu, D. F.; Kristal, B. S.; Hao, Y.; Schweikert, C. Cognitive diversity: A measurement of dissimilarity between multiple scoring systems. *Journal of Interconnection Networks* **2019**, *19*, 1940001.

(22) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548* **2021**,

(23) Hurley, L.; Kristal, B. S.; Sirimulla, S.; Schweikert, C.; Hsu, D. F. Multi-Layer Combinatorial Fusion Using Cognitive Diversity. *IEEE Access* **2020**, *9*, 3919–3935.

(24) Hsu, D.; Shapiro, J.; Taksa, I. Methods of data fusion in information retreival: Rank vs. Score combination. *DIMACS Technical Report* **2002**, *58*, 662–667.

(25) Mitra, A.; Ghosh, S.; Mohapatra, A.; Chakrabarti, S. Appliance Identification via Combinatorial Fusion Analysis-Assisted Bayesian-Optimized Classifier. *IEEE Transactions on Smart Grid* **2023**,

(26) Tang, Y.; Li, Z.; Nellikkal, M. A. N.; Eramian, H.; Chan, E. M.; Norquist, A. J.; Hsu, D. F.; Schrier, J. Improving data and prediction quality of high-throughput perovskite synthesis with model fusion. *Journal of Chemical Information and Modeling* **2021**, *61*, 1593–1602.

(27) Zhang, Z.; Schweikert, C.; Shimojo, S.; Hsu, D. Improving prediction quality of face image preference using combinatorial fusion algorithm. *Brain Informatics* **2023**,

(28) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the

22nd ACM SIGDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794.

(29) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.

(30) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.

(31) Schapire, R. E. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*; Springer, 2013; pp 37–52.

(32) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.

(33) Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine Learning Methods for Small Data Challenges in Molecular Science. *Chemical Reviews* **July 12, 2023**,