Journal of Cheminformatics

RESEARCH Open Access

CLAIRE: a contrastive learning-based predictor for EC number of chemical reactions

Zishuo Zeng[1*], Jin Guo[1], Jiao Jin[1] and Xiaozhou Luo[2*]

Abstract

Predicting EC numbers for chemical reactions enables efficient enzymatic annotations for computer-aided synthesis planning. However, conventional machine learning approaches encounter challenges due to data scarcity and class imbalance. Here, we introduce CLAIRE (Contrastive Learning-based Annotation for Reaction's EC), a novel framework leveraging contrastive learning, pre-trained language model-based reaction embeddings, and data augmentation to address these limitations. CLAIRE achieved notable performance improvements, demonstrating weighted average F1 scores of 0.861 and 0.911 on the testing set (n = 18,816) and an independent dataset (n = 1040) derived from yeast's metabolic model, respectively. Remarkably, CLAIRE significantly outperformed the state-of-the-art model by 3.65 folds and 1.18 folds, respectively. Its high accuracy positions CLAIRE as a promising tool for retrosynthesis planning, drug fate prediction, and synthetic biology applications. CLAIRE is freely available on GitHub (https://github.com/zishuozeng/CLAIRE).

Scientific contribution

This work employed contrastive learning for predicting enzymatic reaction's EC numbers, overcoming the challenges in data scarcity and imbalance. The new model achieves the state-of-the-art performance and may facilitate the computer-aided synthesis planning.

Keywords Reaction EC number, Contrastive learning, Reaction embeddings, Metabolic model, Computer-aided synthesis planning

Introduction

Enzymes are pivotal in catalyzing biochemical reactions vital for life processes. Central to the classification and nomenclature of enzymes is the Enzyme Commission (EC) number system, providing a systematic framework for organizing and understanding enzymatic activities. The EC numbers are formatted into four hierarchical levels. According to the IUBMB Enzyme Nomenclature [1], the first level denotes the most basic classification of enzyme functions, including oxidoreductase (EC 1), transferase (EC 2), hydrolase (EC 3), lyase (EC 4), isomerase (EC 5), ligase (EC 6), and translocase (EC 7). The following second level designates the group or bond where enzyme acts upon, e.g., EC 2.3 indicates acyltransferase under the transferase group (EC 2). The third level specifies the enzymatic reaction, e.g., EC 2.3.2 refers to aminoacyltransferase. The last component is a serial number assigned to the enzyme with specific substrate [2] in sequential order [3], e.g., EC 2.3.2.8 is assigned to arginyltransferase.

The fast and automated annotation of EC number for protein sequences has become especially crucial since the

*Correspondence:
Zishuo Zeng
zengzishuo@synceres.com
Xiaozhou Luo
xz.luo@siat.ac.cn
[1] Synceres Biosciences Co. Ltd., Shenzhen 518100, China
[2] Shenzhen Key Laboratory for the Intelligent Microbial Manufacturing of Medicines, Key Laboratory of Quantitative Synthetic Biology, Center for Synthetic Biochemistry, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

emergence of next generation sequence, which results in rapid accumulation of biological sequences with unknown functions [4]. Many efforts have been made aiming to achieve this goal, ranging from traditional bioinformatics methods [5] to machine learning-based methods [6–9]. However, accurate prediction of EC numbers for proteins is not straightforward, because the total types of complete EC numbers (four levels) are enormous (over 5000 [3]) and the distribution of the available sequences per EC number is highly imbalanced—some EC numbers have tens of thousands of affiliated protein sequences whereas some may only have a handful. Consequently, no highly reliable EC number predictor is in place until CLEAN [10] being introduced very recently. CLEAN adopted a pre-trained large language model for effective protein feature extraction and contrastive learning as the model architecture to overcome the limitations of data imbalance, dubbing it the state-of-the-art method for protein EC number prediction.

While EC numbers are commonly used for annotating enzymes, their application extends to annotating the corresponding reactions catalyzed by these enzymes. However, existing tools for predicting EC-reaction relationships suffer from unsatisfactory performance due to several reasons. First, unlike protein sequence, the EC number of a chemical reaction can be manually determined by experts, which is more reliable than a computational tool. Second, the available EC-reaction datasets are much smaller compared to those for EC-enzyme annotations. For instance, the Rhea database [11] only contains about 21 thousand EC-reaction entries, whereas UniProt [12] contains > 250 thousands reviewed EC-enzyme sequence entries.

Although a reaction's EC number can be manually determined by experts, automated prediction is particularly crucial in the era of synthesis biology, where knowledge of synthesis reactions and metabolic pathways is essential for producing desired compounds in microbial factories [13]. To this end, many tools have been developed for computer-aided synthesis planning (CASP) [14]. CASP may generate large amount of candidate reactions, necessitating automated EC number annotation due to the impracticality of manual curation. With the aid of a reliable EC-reaction predictor, these candidate reactions can be annotated with EC numbers. Meanwhile, the protein sequences derived from transcriptomic or proteomic data can also be annotated with EC numbers by tools such as CLEAN [10]. The mutual presence of the same EC numbers on protein level and reaction level can serve as cross-validation, which facilitates the enzyme mining process for the desired reactions. Altogether, an efficient predictor of EC-reaction predictor is pivotal for enzyme mining and synthesis biology overall.

Most existing reaction-EC annotation tools primarily rely on similarity search against annotated reactions [3, 15–17]. Reaction similarity can be constructed by fingerprints [15], atom–atom mapping [3], types of bonds [16], and mutual information [18]. The similarity-based methodology assumes that the inter-reaction similarity can well capture the complex relationships between reaction similarity and EC labels. However, this assumption often fails especially when large molecules experience relatively minor local changes or when large cofactors participate in the reaction [17]. Matsuta et al. attempted to use machine learning to capture the nonlinear relationship between reaction similarity and EC labels [18]. But they treated the EC number prediction task as a binary classification task, which does not perform well in multi-class tasks [19]. Most recently, a deep learning-based multi-class model called Theia [20] was developed for reaction-EC prediction. Although Theia was trained on a dataset larger than previous ones, the data imbalance problem was still not particularly addressed, which motivated us to develop a new model with higher performance.

Inspired by CLEAN [10], we developed a novel reaction-EC number predictor using contrastive learning architecture, which is demonstrated to be beneficial in remedying data imbalance in classification tasks [21–23]. The data input to the contrastive learning network includes differential reaction fingerprints (DRFP) [24] and embeddings derived from a pre-trained language model [25], both of which are shown to be superior to structure-based fingerprints [24]. Moreover, we performed data augmentation by shuffling the order within reactants and within products simultaneously to improve the model robustness. We name this new model CLAIRE (Contrastive Learning-based AnnotatIon for ReactI on's EC). As a result, CLAIRE outperforms Theia substantially, demonstrating its utility in reaction-EC prediction and its promise in enzyme mining for synthetic biology.

Methods

Data curation and processing

Rhea [11] is a commonly used database dedicated for reaction and enzyme mapping, which comprises only about 21 thousand reaction-enzyme pairs with annotated EC numbers. Recently, a much larger reaction-EC dataset called ECREACT [26] has been collected by combining data from Rhea, BRENDA [27], PathBank [28], and MetaNetX [29]. We downloaded the ECREACT dataset [26] (n = 62,222), including 277 3rd-level EC numbers, wherein we removed 101 EC numbers that have less than 10 reaction entries, resulting in 61,817 EC-reaction entries. These remaining EC-reaction entries cover seven 1st-level EC numbers (e.g., 5.-.-.-), 63 2nd-level EC numbers (e.g., 2.1.-.-), and 175 3rd-level EC numbers

(e.g., 1.2.1.-). For each of the 3rd-level EC numbers, we left 10% of samples (reactions) for final testing. For the remaining 90% of samples, we further split the data in 1:9 ratio for validation and training purposes.

To ensure the robustness of the model, we performed data augmentation by shuffling participants within reactants and products. For example, reaction A + B = C + D can be augmented to four reactions: A + B = C + D, B + A = C + D, A + B = D + C, and B + A = D + C. To demonstrate the necessity of the data augmentation, we randomly selected 1000 reactions to perform data augmentation and computed the Euclidean distance of rxnfp embeddings between the original reaction and shuffled reactions (e.g., A + B = C + D and B + A = C + D). We then compared these same-but-shuffled reactions' distances with the distances between actually different reactions (pairwise distances among the 1000 reactions). Result showed substantial overlap–there are 11.7% same-but-shuffled distances that are greater than the 10-percentile of distances between different reactions (Supplementary Fig. 1), confirming the necessity the data augmentation. The data augmentation process resulted in a three-fold size increase in training set (n = 150,226), validation set (n = 16,692), and testing set (n = 18,816).

Feature engineering

We computed embeddings for reactions in our dataset using the rxnfp [25] pre-trained model and differential reaction fingerprints (DRFP) [24]. The rxnfp model is a transformer-based classifier that predicts a reaction's category (e.g., nitro to amino). The rxnfp model was trained on ~ 3 million reactions from Pistachio database (https://www.nextmovesoftware.com/pistachio.html) and UPSTO 50K dataset [30] to predict the reaction category (e.g., nitro to amino). In this study, we fed the enzymatic reactions in SMILES format to rxnfp, from which the resulting model hidden layer is extracted as the embeddings. These embeddings serve as the features to describe the properties of a chemical reaction for our machine learning task. The rxnfp-derived embeddings enable mapping the reaction space properly (i.e., reactions belonging to the same category are placed closer together in the embeddings dimensions) [25] and have been found useful in a variety of downstream machine learning tasks, such as prediction of chemical yields [31], biocatalytic synthesis planning [26], and reaction classification [24].

DRFP, unlike rxnfp, is not a machine learning-based method. It converts a reaction SMILES to a binary fingerprint by comparing the symmetric difference of two sets of circular n-grams extracted from the molecules positioned to the left and right of the reaction arrow [24]. Since the reactions are in SMILES (simplified molecular-input line-entry system) format [32], the reaction embeddings can be readily applied. Both DRFP and rxnfp converts a SMILES-based reaction to a 256-long vector, resulting in a final feature set with 512 numeric values long.

Curation of yeast metabolic model data

We curated an additional large-batch dataset from yeast's metabolic model for further validation on CLAIRE. We first obtained the gene-reaction mapping relationships from the yeast metabolic model, iMM904 [33], where genes are listed as IDs (e.g., YOR190W) and reaction comprises of model-specific IDs for metabolites (e.g., 2hp6mbq_m). To obtain the ground truths of EC number for each gene-reaction pair, we obtained the sequences of the yeast gene IDs and BLASTed [34] them against the UniProt database [12], from which we acquired the EC number annotations. To enable the reaction-EC prediction by CLAIRE, metabolites of reactions in iMM904 need to be converted to SMILES format. Thus, we first mapped the metabolite IDs to metabolite names through the inherent mappings available in iMM904; we then mapped metabolite names to SMILES through a mapping table downloaded from ChEBI database [35], which is a comprehensive database for chemicals with information of chemical names, synonyms, canonical SMILES, etc. Note that some metabolite names in iMM904 are are difficult to be standardized using regular expression. Thereby, we have to manually fix the metabolite names when necessary, for examples, "2 Hydroxy hexadecanal C16H32O2" to "2-hydroxyhexadecanal", "Nicotinamide adenine dinucleotide phosphate—reduced" to "NADPH". Also, there are many metabolite names that failed to be identified based on the ChEBI database and therefore abandoned, such as "Mannose inositol P 2 ceramide ceramide 1 26C", "Peptide C2H4NO2RC2H2NOR", and "TRNA(Phe)". As a result, we collected 1122 reaction-EC (genes) pairs with 617 unique genes and 578 unique reactions. After removing duplicates on genes, we obtained 456 reaction-EC pairs as the positive set.

For an objective evaluation, we also compiled a negative set. Since there could be tremendous possible false reactions for negative reaction-EC pairing, we selected the reaction that is closest to the reaction in the positive reaction-EC pair, i.e., the false reaction only differs on the third level from the positive reaction-EC pair. This way, we can significantly narrow down the candidate pool and maximize the challenge of the negative set to the model. Note that, since some reactions in the positive set may correspond to multiple EC numbers, we first expanded the reaction-EC(s) to single reaction-EC pairs then performed the negative set construction. As a result, the negative set (n = 584) is slightly larger than the positive set.

Model implementation

We trained an individual model for 1st-level, 2nd-level, and 3rd-level EC number prediction, respectively. The methodology and code for training and prediction were modified based on CLEAN's [10] framework (https://github.com/tttianhao/CLEAN).

We adopted the Triplet Margin Loss (TML) [36] as the training strategy, where the sampling procedures are as the following:

1. A selected data point is set as Anchor;
2. A data point with the same label (EC number) as the Anchor is randomly selected to form a Positive pair;
3. A data point with a different label from the Anchor is randomly selected as a Negative pair.

Such an Anchor-Positive–Negative combination constitutes a triplet. The objective is to optimize the model parameters by the TML loss function (Eq. 1), i.e., to minimize Anchor-Positive distance and maximize the Anchor-Negative distance, respectively, while maintaining a distance greater than a predefined margin, thereby allowing the network to better learn the differences between samples. Each sample in the entire dataset is exhaustively iterated as Anchor.

$$TML = \sum_{i}^{N} [\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha]_+ \quad (1)$$

where $[\cdot]_+ = \begin{cases} 0, x < 0 \\ \cdot, x \geq 0 \end{cases}$

Here, the features of Anchor, Positive, and Negative are represented by $f(x_i^a)$, $f(x_i^p)$, and $f(x_i^n)$, respectively. α represents the predefined margin between Positive and Negative pairs.

Apart from the TML strategy [36], CLEAN also proposed Supcon Hard Loss (SHL) [10, 37]. SHL aims to challenge the model by selecting negative samples that is similar to the Anchor. However, in our dataset, there exists cases where the feature distances are small whereas the labels are very different (i.e., differ by the first level). Also, SHL limits the options of Negative, sacrificing the diversity during triplet construction. These limitations impede the learning efficiency and subject the model to overfitting, and thus the SHL strategy is not adopted.

Hyperparameter tuning

For the TML strategy, the hyperparameter α may have an impact on the model performance and training efficiency, we therefore perform experiments to find the optimal α from 0.5, 1, 1.5, and 2 with a default set of hyperparameters (number of layers = 5; learning rate = 0.0001; hidden dimension = 1280). Results showed that the model converges most quickly and both training and validation losses are minimal at α=1 (Supplementary Fig. 2). For the hyperparameters of the fully connected neural network, we first performed grid search for number of layers and hidden dimension. After finding the optimal set of these two hyperparameters, we continued to optimize the learning rate. The final set of hyperparameters are as follows: number of layers = 5; learning rate = 0.0001; hidden dimension = 1280. Code was implemented in Python (v 3.10.4) with PyTorch package (v 1.11.0) and scikit-learn (v 1.2.0).

Ablation study

Since we used two embedding strategies (rxnfp, DRFP) for data featurization, we performed an ablation study to investigate whether combining the two types of embeddings is better than either one alone. We trained the model on training set and evaluated the WAF1 on validation set using the optimized set of hyperparameters. Result showed that combining rxnfp and DRFP is slightly better than using only one embeddings strategy (Supplementary Fig. 3).

Prediction

The EC number prediction for a given reaction is made by the following: (1) compare the Euclidean distances between the input reaction's last model layer to the last model layer for each of the reactions with known EC numbers; (2) assign the EC number whose reaction has the closest distance to the input reaction; (3) if desired, output the top K EC numbers ranked by the distances.

Evaluation metric

To properly evaluate the model performance on the testing sets, we used weighted average F1 score (WAF1; Eq. 2), which allows adjustment of the final metric by sample size, ensuring a fairer evaluation.

$$\text{WAF1} = \frac{\sum_{i=1}^{C} N_i \times F1_i}{\sum_{i=1}^{C} N_i} \quad (2)$$

where C, $F1_i$ and N_i are number of EC numbers, the F1 score of the i-th EC number, and the sample size of the i-th EC number, respectively. F1 score for a given EC number can be calculated by Eq. 3.

$$F1score = \frac{2tp}{2tp + fp + fn} \quad (3)$$

where tp, fp, and fn are true positive, false positive, and false negative, respectively.

Since three models are developed separately for three levels of EC numbers, it is desired to evaluate whether the three models agree with each other on the same reaction.

We therefore proposed *consistency* (Eq. 4) to measure the level of consensus by the three models.

$$consistency = \frac{\sum_{i=1}^{N} I_i}{N} \times 100\% \quad (4)$$

$$I_i = \begin{cases} 1, \ if \ a1 = b1 = c1 \ and \ b2 = c2 \\ 0, \ else \end{cases}$$

where N is number of sample size; the 1st-level, 2nd-level, and 3rd-level predictions are "a1", "b1.b2", and "c1.c2.c3", respectively.

Comparison with other model

We compared our model with the state-of-the-art model, Theia [20], on the testing set and the yeast metabolic model dataset. Theia can be obtained from https://github.com/daenuprobst/theia. Note that Theia includes two sets of model: Theia-RHEA and Theia-ECREACT, which were trained on the Rhea dataset and ECREACT dataset, respectively. We incorporated both sets of models for comparison.

Results

Data curation

We obtained and cleaned the reaction-EC number dataset from ECREACT [26], totaling 61,817 unique reaction entries. Since the 4th-level is a serial number and does not confer learnable information [20], we limit our reaction-EC prediction task to the 3rd-level, which is in line with other existing predictors [15, 16, 20]. For each of the 3rd-level EC numbers, we separated 10% reaction-EC entries for final testing; we further separated the remaining 90% data in 1:9 ratio for the split of validation set and training set. Since a reaction may have various order among reactants and products, the arrangement of reactants or products' order may influence the model prediction. In particular, we observed that a shuffled reaction could have very different rxnfp [25] embeddings than the original one (Methods; Supplementary Fig. 1), suggesting that changing reaction participants' order may substantially change the features of a sample. We therefore performed data augmentation by shuffling participants within reactants and products (Methods) to a) ensure the robustness of model prediction in terms of reaction participants' order; and b) for data augmentation purpose. The data augmentation results in a three-fold size increase in training set (n = 150,226), validation set (n = 16,692), and testing set (n = 18,816).

Model implementation

We adopted contrastive learning [37] as our model architecture for the reaction-EC prediction. We trained an individual model for 1st-level, 2nd-level, and 3rd-level EC number prediction, respectively. Contrastive learning (Fig. 1) operates on the principle of maximizing the similarity between positive pairs (reactions belonging to the same EC number) while minimizing the similarity between negative pairs (reactions belonging to different EC numbers). We used the rxnfp [25] pre-trained model and the DRFP framework [24] for feature extraction (Methods). Both rxnfp and DRFP have been shown to be effective feature extraction techniques in a variety of reaction-related downstream tasks [31, 38, 39]. In our experiments, we observed that combining rxnfp and DRFP achieves higher WAF1 (0.942) on training set than rxnfp (0.895) or DRFP (0.940) alone.

Performance comparison

We compared the performance of CLAIRE and Theia on the augmented testing set (n = 18,816). As a result, the weighted average F1 score (WAF1; Eq. 2) of CLAIRE for 1st-level, 2nd-level, and 3rd-level EC number prediction are much higher (by 0.574–0.639) than Theia (Fig. 2). Besides, we defined *consistency* (Eq. 4) to measure the agreement of the three levels of predictions. For example, if the three models predict a reaction to be 2.-.-.-, 2.1.-.-, and 2.1.3.-, respectively, then it is considered *consistent*; otherwise (e.g., 2.-.-.-, 2.4.-.-, and EC 2.1.3.-) is not. We then computed and compared the *consistency* of Theia and CLAIRE on testing set. Results show that 82.01% of the reactions on testing set are *consistent*, while only 15.52% of the reactions received consistent predictions from Theia. These results collectively suggest that CLAIRE is more accurate and less self-contradictory. In addition, we observed that 97% of the non-augmented samples in the testing set have consistent predictions when augmented by shuffling reactants or product's order, highlighting that the model is robust against varying reaction participants' input order.

Next, we managed to curate another large-size reaction-EC dataset for further validation on CLAIRE (Fig. 3A; Methods). We extracted a total of 456 reaction-gene pairs as positive set from the yeast metabolic model, iMM904 [33]. After annotation and processing (Methods), we obtained a dataset containing 456 reaction-EC number entries as the positive set. For the negative set, we compiled 584 false enzyme-reaction pairs by substituting the reaction in a true enzyme-reaction pair with the most similar analog from the reaction pool, so that negative data are more challenging to the model. The final dataset has 1040 entries (456 positive

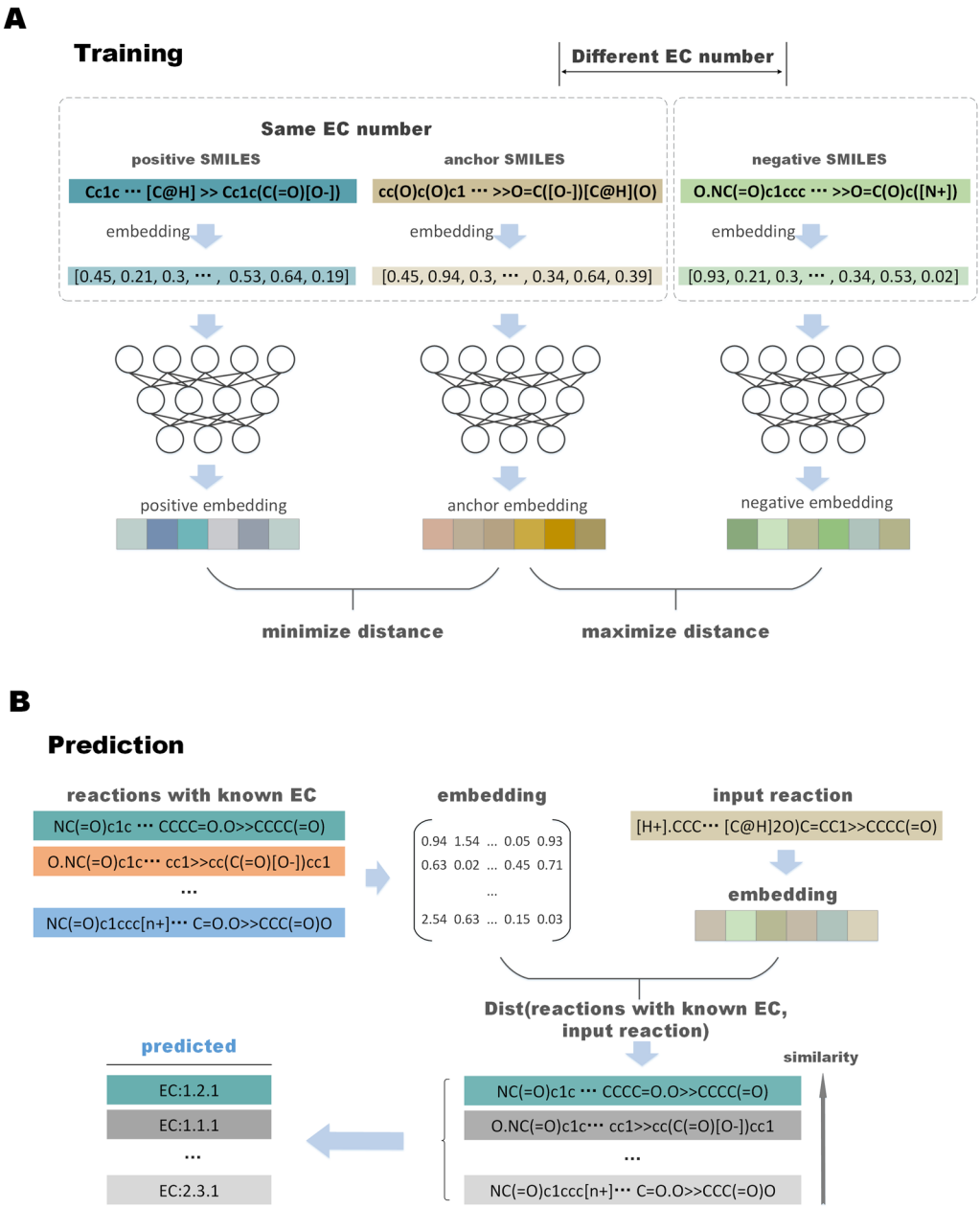

Fig. 1 Contrastive learning framework in this study. **A** shows the principle of training CLAIRE: (1) randomly selecting a sample as Anchor; (2) randomly selecting a Positive sample (same EC number as Anchor's) and a Negative sample (different EC number as Anchor's); (3) optimize the neural network by minimizing the Anchor-Positive distance while maximizing the Anchor-Negative distance. **B** shows how the prediction is made for a query reaction: (1) calculate Euclidean distance in terms of model last layer vector between the query and each of the reference reactions with known EC numbers; (2) identify the closest reference reaction and output its EC number as the prediction

and 584 positive). For evaluation purpose, we calculated the top 1–3 WAF1 on the positive and negative data (Methods). As a result, CLAIRE's top 1–3 WAF1 are 0.911, 0.93, and 0.926 (Fig. 3), respectively, outperforming Theia (0.437, 0.450, and 0.455, respectively). This result once again demonstrated the superior performance of CLAIRE.

Discussion

The purpose of developing a reaction-EC number predictor lies in making high-throughput predictions, instead of in small batches. This is because, in reality, small batches of reactions can be directly annotated by experts, whereas annotating large batches of reactions manually is practically infeasible. With that said, it is not so meaningful to

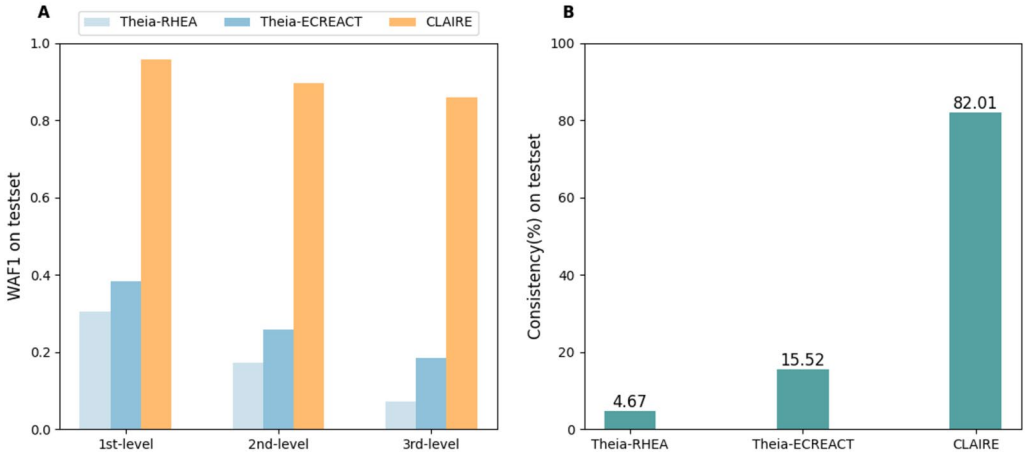

Fig. 2 Performance comparison between CLAIRE and Theia (including Theia-RHEA and Theia-ECREACT) on testing set. **A** shows the WAF1 (weighted average F1 score) of the models on the testing set. Since both CLAIRE and Theia offers a model on three EC number levels, the WAF1 on three levels are shown separately. **B** shows the consistency of CLAIRE and Theia on the testing set

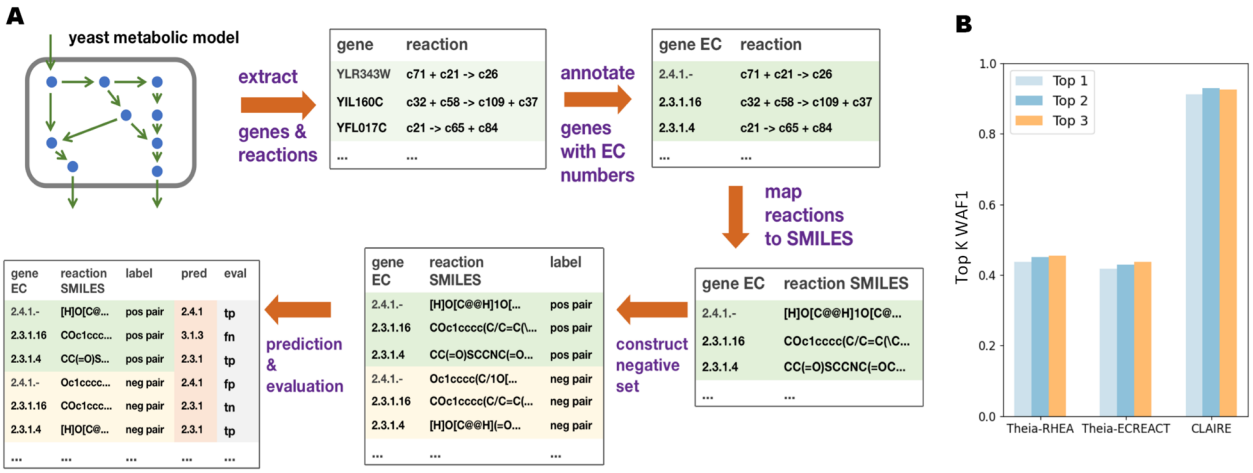

Fig. 3 Data curation of the yeast metabolic model dataset and evaluation result. **A** shows the procedures of curating data from the yeast metabolic model: (1) extracting gene IDs and reactions catalogued in the yeast metabolic model (iMM904); (2) annotate the genes with EC numbers and map the reaction compounds to SMILES, these gene-derived EC numbers and the corresponding reactions serve as the positive pairs; (3) construct negative pairs by substituting the reaction in a positive EC-reaction pair with a close analog (see Methods); (4) apply the predictors on the curated dataset (positive set and negative set) and calculate the top 1, top 2 and top 3 WAF1 (shown in **B**)

further test CLAIRE on an additional small dataset on top of our testing sets. Meanwhile, we still want to evaluate the utility of CLAIRE in a larger extent. The enzyme-reaction pairs catalogued in metabolic models serve for this purpose. Metabolic models are quantitative representations of metabolic reactions with associated genes, enzymes, metabolites, and compartmentalization in an organism or cells [40]. It can be used for various tasks such as flux balance analysis to guide metabolic engineering [41]. In this study, we showed that CLAIRE can accurately distinguish positives (true enzyme-reaction pairs) and negatives (false enzyme-reaction pairs) derived from the yeast metabolic model. This success further highlights the promising utility of CLAIRE in a high-throughput setting.

High-throughput prediction of reaction EC numbers is particularly valuable in applications like retrosynthesis planning [42] and drug fate prediction [43]. Retrosynthesis prediction refers to predicting the reactants leading to a product of interest [44]. Retrosynthesis planning involves identifying the reactants required to produce a target compound, often through multiple iterative steps.

This process generates a reaction network that traces the possible pathways from raw materials to the desired products [45]. Each step along the proposed pathway may yield multiple potential intermediate reactants, resulting in a large and complex network of reactions. By incorporating CLAIRE-predicted EC numbers into this network, relevant enzymes can be assigned to reactions, thereby increasing the feasibility of achieving the desired synthesis. However, CLAIRE's utility in retrosynthesis is limited to biocatalytic processes, as it assigns EC numbers regardless of whether the reaction can be enzymatically catalyzed [46, 47], as CLAIRE would annotate an EC number for any query reaction, regardless of whether the reaction can be catalyzed by an enzyme or not. Recent advances in deep learning-based bio-retrosynthesis tools [46] provide an ideal framework for integrating CLAIRE's predictions. Another promising application of CLAIRE is in drug fate prediction, which models the metabolic transformations and pathways a drug compound undergoes in the human body [43]. This process, which can be viewed as the reverse of retrosynthesis planning, involves predicting successive reactions that metabolize the drug. Similar to retrosynthesis, the stepwise predictions often generate large networks of potential reactions, as each reactant can produce multiple products [48]. CLAIRE's ability to annotate these reactions with EC numbers facilitates various analyses, such as evaluating drug toxicity or identifying key metabolic pathways [49].

The primary limitation of CLAIRE is its incomplete coverage of three-level EC numbers, which stems from a lack of sufficient data. Overcoming this challenge is inherently difficult without the availability of more training data. Additionally, many enzymatic reactions remain unexplored and uncharacterized, leaving their corresponding EC numbers undefined. However, the utility of CLAIRE is not highly sensitive to the pace of new EC number discovery. This is because the addition of new three-level EC numbers has been relatively slow; our analysis revealed that only 17 new three-level EC numbers have been incorporated into reviewed entries in UniProt over the past decade. We provide a list of the EC numbers covered by CLAIRE (Supplementary Table 1), where users should check whether the EC numbers of users' interest (if any) are in that list before applying CLAIRE.

Conclusion

In summary, we built CLAIRE for reaction-EC number prediction leveraging contrastive learning and data augmentation to overcome limited data size and data imbalance. We demonstrated that CLAIRE outperforms the state-of-the-art model in terms of accuracy and consistency, suggesting that CLAIRE may facilitate tasks in computer-aided synthesis planning, such as retrosynthesis planning and drug fate prediction.

Abbreviations

EC	Enzyme commission
CASP	Computer-aided synthesis planning
DRFP	Differential reaction fingerprints
WAF1	Weighted average F1 score

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00944-8.

Additional file 1.

Additional file 2.

Acknowledgements

We would like to thank all researchers who made the data and tools relevant to this study available. We also thank all the funding agencies for supporting this research.

Author contributions

Z.Z. conceived the ideas, curated the data, and wrote the manuscript; J.G. implemented the model and performed visualization; J.J. obtained EC numbers for the yeast dataset; X.L. reviewed and revised the manuscript.

Funding

The Project Supported by National Key R&D Program of China (2019YFA0904100), National Natural Science Foundation of China (32071421), Guangdong Basic and Applied Basic Research Foundation (2021B1515020049), Shenzhen Science and Technology Program (ZDSYS20210623091810032, KJZD20230923115906013, KJZD20230923115901003 and RCYX20200714114736026), SIAT Distinguished Young Scholars (E4G021).

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Competing interests

Xiaozhou Luo has a financial interest in Demetrix and Synceres.

Received: 23 July 2024 Accepted: 19 December 2024
Published online: 07 January 2025

References

1. Tipton K, Boyce S (2000) History of the enzyme nomenclature system. Bioinformatics 16(1):34–40
2. McDonald AG, Tipton KF (2023) Enzyme nomenclature and classification: the state of the art. FEBS J 290(9):2214–2231
3. Rahman SA, Cuesta SM, Furnham N, Holliday GL, Thornton JM (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat Methods 11(2):171–174
4. Zaru R, Magrane M, Orchard S, Uniprot Consortium (2020) Challenges in the annotation of pseudoenzymes in databases: the UniProtKB approach. FEBS J 287(19):4114–4127
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
6. Ryu JY, Kim HU, Lee SY (2019) Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci 116(28):13996–14001
7. Sanderson T, Bileschi ML, Belanger D, Colwell LJ (2023) ProteInfer, deep neural networks for protein functional inference. Elife 12:e80942

8. Dalkiran A, Rifaioglu AS, Martin MJ, Cetin-Atalay R, Atalay V, Doğan T (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. BMC Bioinformatics 19:1–13
9. Li Y, Wang S, Umarov R, Xie B, Fan M, Li L et al (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics 34(5):760–769
10. Yu T, Cui H, Li JC, Luo Y, Jiang G, Zhao H (2023) Enzyme function prediction using contrastive learning. Science 379(6639):1358–1363
11. Bansal P, Morgat A, Axelsen KB, Muthukrishnan V, Coudert E, Aimo L et al (2022) Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res 50(D1):D693–D700
12. Uniprot Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47(D1):D506–D515
13. Dudley QM, Karim AS, Jewett MC (2015) Cell-free metabolic engineering: Biomanufacturing beyond the cell. Biotechnol J 10(1):69–82
14. Cook A, Johnson AP, Law J, Mirzazadeh M, Ravitz O, Simon A (2012) Computer-aided synthesis design: 40 years on. Wiley Interdiscip Rev Comput Mol Sci 2(1):79–107
15. Hu Q-N, Zhu H, Li X, Zhang M, Deng Z, Yang X et al (2012) Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. PLoS ONE 7(12):e52901
16. Yamanishi Y, Hattori M, Kotera M, Goto S, Kanehisa M (2009) E-zyme: predicting potential EC numbers from the chemical transformation pattern of substrate-product pairs. Bioinformatics 25(12):i179–i186
17. Hadadi N, MohammadiPeyhani H, Miskovic L, Seijo M, Hatzimanikatis V (2019) Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. Proc Natl Acad Sci 116(15):7298–7307
18. Matsuta Y, Ito M, Tohsato Y (2013) ECOH: an enzyme commission number predictor using mutual information and a support vector machine. Bioinformatics 29(3):365–372
19. Mudiyanselage DLBAG. Multi-Label Classification Using Higher-Order Label Clusters. University of Nebraska at Omaha. 2018.
20. Probst D (2023) Explainable prediction of catalysing enzymes from reactions using multilayer perceptrons. bioRxiv. https://doi.org/10.1101/2023.01.28.526009
21. Zhang J, Zou J, Su Z, Tang J, Kang Y, Xu H et al (2022) A class-aware supervised contrastive learning framework for imbalanced fault diagnosis. Knowl-Based Syst 252:109437
22. Marrakchi Y, Makansi O, Brox T (2021) Fighting class imbalance with contrastive learning. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp 466–476
23. Zhou J, Li G, Wang R, Chen R, Luo S (2023) A novel contrastive self-supervised learning framework for solving data imbalance in solder joint defect detection. Entropy 25(2):268
24. Probst D, Schwaller P, Reymond J-L (2022) Reaction classification and yield prediction using the differential reaction fingerprint DRFP. Digital Discovery 1(2):91–97
25. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T et al (2021) Mapping the space of chemical reactions using attention-based neural networks. Nat Mach Intell 3(2):144–152
26. Probst D, Manica M, Teukam Y, Castrogiovanni A, Paratore F, Laino T (2022) Biocatalysed synthesis planning using data-driven learning. Nat Commun 13(1):964
27. Schomburg I, Chang A, Schomburg D (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30(1):47–49
28. Wishart DS, Li C, Marcu A, Badran H, Pon A, Budinski Z et al (2020) PathBank: a comprehensive pathway database for model organisms. Nucleic Acids Res 48(D1):D470–D478
29. Ganter M, Bernard T, Moretti S, Stelling J, Pagni M (2013) MetaNetX org a website and repository for accessing, analysing and manipulating metabolic networks. Bioinformatics 29(6):815–816
30. Schneider N, Stiefl N, Landrum GA (2016) What's what: The (nearly) definitive guide to reaction role assignment. J Chem Inf Model 56(12):2336–2346
31. Schwaller P, Vaucher AC, Laino T, Reymond J-L (2021) Prediction of chemical reaction yields using deep learning. Mach Learn Sci Technol 2(1):015016
32. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Che Inform Comput Sci 28(1):31–36
33. Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M et al (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 26(10):1155–1160
34. Pundir S, Martin MJ, O'Donovan C, UniProt Consortium (2016) UniProt tools. Curr Protocols Bioinform 53(1):1.29
35. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, Mcaught A et al (2007) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36(suppl_1):D344–D350
36. Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; 815–23.
37. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P et al (2020) Supervised contrastive learning. Adv Neural Inf Process Syst 33:18661–18673
38. Kroll A, Rousset Y, Hu X-P, Liebrand NA, Lercher MJ (2023) Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. Nat Commun 14(1):4139
39. Schleinitz J, Langevin M, Smail Y, Wehnert B, Grimaud L, Vuilleumier R (2022) Machine learning yield prediction from NiCOlit, a small-size literature data set of nickel catalyzed C-O couplings. J Am Chem Soc 144(32):14722–14730
40. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY (2019) Current status and applications of genome-scale metabolic models. Genome Biol 20:1–18
41. Raman K, Chandra N (2009) Flux balance analysis of biological systems: applications and challenges. Brief Bioinform 10(4):435–449
42. Sun Y, Sahinidis NV (2022) Computer-aided retrosynthetic design: fundamentals, tools, and outlook. Curr Opin Chem Eng 35:100721
43. Kazmi SR, Jun R, Yu M-S, Jung C, Na D (2019) In silico approaches and tools for the prediction of drug metabolism and fate: a review. Comput Biol Med 106:54–64
44. Zhong Z, Song J, Feng Z, Liu T, Jia L, Yao S et al (2024) Recent advances in deep learning for retrosynthesis. Wiley Interdiscip Rev Comput Mol Sci 14(1):e1694
45. Schwaller P, Vaucher AC, Laplaza R, Bunne C, Krause A, Corminboeuf C et al (2022) Machine intelligence for chemical reaction space. Wiley Interdiscip Rev Comput Mol Sci 12(5):e1604
46. Zheng S, Zeng T, Li C, Chen B, Coley CW, Yang Y et al (2022) Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. Nat Commun 13(1):3342
47. Kim T, Lee S, Kwak Y, Choi MS, Park J, Hwang SJ et al (2024) READRetro: natural product biosynthesis predicting with retrieval-augmented dual-view retrosynthesis. New Phytol 243(6):2512–2527
48. Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, Greiner R, Manach C, Wishart DS (2019) BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. J Cheminform 11:1–25
49. Kirchmair J, Williamson MJ, Tyzack JD, Tan L, Bond PJ, Bender A et al (2012) Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. J Chem Inf Model 52(3):617–648

Publisher's Note