## A Proof of Theorem 1

**Definition 1** (Classical Learning Problem ($\mathcal{CLP}$))**.** *Given a knowledge base $\mathcal{K}$, a target concept $T$, a set of positive examples $E^+ = \{e_1^+, e_2^+, \ldots, e_{n_1}^+\}$, and a set of negative examples $E^- = \{e_1^-, e_2^-, \ldots, e_{n_2}^-\}$, the learning problem is to find a class expression $C$ such that $T$ does not occur in $C$ and for $\mathcal{K}'_C = \mathcal{K} \cup \{T \equiv C\}$, we have that $\mathcal{K}'_C \models C(E^+)$ and $\mathcal{K}'_C \not\models C(E^-)$.*

**Definition 2** (Generalized Learning Problem ($\mathcal{GLP}$))**.** *Given a knowledge base $\mathcal{K}$, a target concept $T$, and sets of positive/negative examples $E^+ = \{e_1^+, e_2^+, \ldots, e_{n_1}^+\}$ and $E^- = \{e_1^-, e_2^-, \ldots, e_{n_2}^-\}$, the learning problem is to find non-empty subsets $\mathcal{E}^+ \subseteq E^+$, $\mathcal{E}^- \subseteq E^-$ with the following properties*

1. $\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}^+, \mathcal{E}^-) \neq \emptyset$
2. $\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}^+, \mathcal{E}^-) \subseteq \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, E^+, E^-)$
3. *There do not exist non-empty subsets $\mathcal{E}'^+ \subseteq E^+$, $\mathcal{E}'^- \subseteq E^-$ such that $|\mathcal{E}'^+| + |\mathcal{E}'^-| < |\mathcal{E}^+| + |\mathcal{E}^-|$ and $\emptyset \neq \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}'^+, \mathcal{E}'^-) \subseteq \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, E^+, E^-)$,*

where $|.|$ denotes the cardinality of a set.

**Theorem 1.** *$\mathcal{GLP}$ has a solution if and only if $\mathcal{CLP}$ has one.*

To prove Theorem 1, we define $\Gamma(\mathcal{E}^+, \mathcal{E}^-, \mathcal{E}'^+, \mathcal{E}'^-)$ to be the following logical expression which we call the $\Gamma$ condition:

$$\begin{aligned} \big[|\mathcal{E}'^+| + |\mathcal{E}'^-| < |\mathcal{E}^+| + |\mathcal{E}^-|\big] \wedge \\ \big[\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}'^+, \mathcal{E}'^-) \neq \emptyset\big] \wedge \\ \big[\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}'^+, \mathcal{E}'^-) \subseteq \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, E^+, E^-)\big] \end{aligned}$$

for any non-empty sets $\mathcal{E}^+, \mathcal{E}^-, \mathcal{E}'^+$, and $\mathcal{E}'^-$. Here, $\wedge$ is the "logical and" operator.

*Proof.* First, assume that $\mathcal{GLP}$ has a solution. Then (by definition), there exist non-empty subsets $\mathcal{E}^+ \subseteq E^+$, $\mathcal{E}^- \subseteq E^-$ such that $\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}^+, \mathcal{E}^-) \neq \emptyset$ and $\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}^+, \mathcal{E}^-) \subseteq \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, E^+, E^-)$ (properties *1.* and *2.*). Let $C$ be an arbitrary element in $\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}^+, \mathcal{E}^-)$. Then, $C \in \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, E^+, E^-)$ and therefore $C$ is a solution to $\mathcal{CLP}$.

It remains to prove that if $\mathcal{CLP}$ has a solution, then $\mathcal{GLP}$ also has one. Assume that $C_0$ is a solution to $\mathcal{CLP}$, and let $\mathcal{E}_0^+ = E^+$ and $\mathcal{E}_0^- = E^-$. Then, $C_0 \in \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}_0^+, \mathcal{E}_0^-)$ and $\mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, \mathcal{E}_0^+, \mathcal{E}_0^-) \subseteq \mathcal{S}_{\mathcal{CLP}}(\mathcal{K}, T, E^+, E^-)$ (i.e., properties *1.* and *2.* are satisfied by $(\mathcal{E}_0^+, \mathcal{E}_0^-)$). If there do not exist subsets $\mathcal{E}_1^+$, $\mathcal{E}_1^-$ such that $\emptyset \neq \mathcal{E}_1^+ \subseteq E^+$, $\emptyset \neq \mathcal{E}_1^- \subseteq E^-$, and $\Gamma(\mathcal{E}_0^+, \mathcal{E}_0^-, \mathcal{E}_1^+, \mathcal{E}_1^-)$ holds (property *3.*), then $(\mathcal{E}_0^+, \mathcal{E}_0^-)$ is a solution to $\mathcal{GLP}$. If such $\mathcal{E}_1^+, \mathcal{E}_1^-$ exist, define $S_n = (\mathcal{E}_n^+, \mathcal{E}_n^-)$ and $\Sigma_n = |\mathcal{E}_n^+| + |\mathcal{E}_n^-|$ for any integer $n \geq 1$ such that $\Gamma(\mathcal{E}_{n-1}^+, \mathcal{E}_{n-1}^-, \mathcal{E}_n^+, \mathcal{E}_n^-)$ holds. Let $S = \{S_n\}_{n \geq 1}$ and $\Sigma = \{\Sigma_n\}_{n \geq 1}$. Then, $S$ and $\Sigma$ are non-empty sets since $(\mathcal{E}_1^+, \mathcal{E}_1^-) \in S$ and $|\mathcal{E}_1^+| + |\mathcal{E}_1^-| \in \Sigma$. The mapping $f : S \to \Sigma : S_n \longmapsto \Sigma_n$ is clearly a bijection (see proof of Lemma 1 below). Moreover, $\Sigma_n$ is a strictly decreasing sequence of integer values due to the fact that for any integers $n_1 > n_2 \geq 1$ such that $S_{n_1}$ and $S_{n_2}$ exist, we have $|\mathcal{E}_{n_1}^+| + |\mathcal{E}_{n_1}^-| < |\mathcal{E}_{n_2}^+| + |\mathcal{E}_{n_2}^-|$ (i.e., $\Sigma_{n_1} < \Sigma_{n_2}$) through the $\Gamma$ condition. Given that $\mathcal{E}_n^+$ and $\mathcal{E}_n^-$ are not empty, we also have $\Sigma_n \geq 2$ for all $n \geq 1$ (at least one positive example and one negative example). The set $\Sigma$ is therefore finite and admits a minimum $\Sigma_{n^*}$. Hence, $f^{-1}(\Sigma_{n^*}) \in S$ is a solution to $\mathcal{GLP}$ as it satisfies all the properties in Definition 2; this completes the proof. □

**Lemma 1.** *The mapping $f : S \to \Sigma : S_n \longmapsto \Sigma_n$ is a bijection.*

*Proof.* We first prove that $f$ is a surjective function. Let $e \in \Sigma$. Then, there exists $n \geq 1$ such that $\Sigma_n = e$ (by definition of $\Sigma$). Consequently, $S_n \in S$ (note the same subscript $n$ as $\Sigma_n$) and we have $f(S_n) = \Sigma_n$. Hence, $f$ is surjective.

We now prove that $f$ is one to one. Let $s_1, s_2 \in S$ such that $s_1 \neq s_2$. By definition of $S$, there exist $n_1 \geq 1$ and $n_2 \geq 1$ such that $s_1 = S_{n_1}$ and $s_2 = S_{n_2}$. Without loss of generality we can assume that $n_1 > n_2$. Then, $|\mathcal{E}_{n_1}^+| + |\mathcal{E}_{n_1}^-| < |\mathcal{E}_{n_2}^+| + |\mathcal{E}_{n_2}^-|$ following the $\Gamma$ condition. In other words, $\Sigma_{n_1} < \Sigma_{n_2}$ and hence $\Sigma_{n_1} \neq \Sigma_{n_2}$. $f$ is therefore one to one. □

## B Hyperparameter Configuration

In Table 1, we report hyperparameter values used in our experiments. We searched for the best values on Carcinogenesis and used them on the rest of the datasets.

| | **Carcino.** | **Mutag.** | **Sem. B.** | **Vicodi** |
|---|---|---|---|---|
| *Epochs* | 400 | 400 | 400 | 400 |
| *Lr* | 0.001 | 0.001 | 0.001 | 0.001 |
| *d* | 50 | 50 | 50 | 50 |
| *Batch_size* | 512 | 512 | 512 | 512 |
| *L* | 48 | 48 | 48 | 48 |
| *gc* | 5 | 5 | 5 | 5 |

Table 1: Hyper-parameter settings per dataset. *Lr* is the learning rate, *d* is the embedding dimension in the embedding model, *L* is the maximum length of expressions synthesized by ROCES, and *gc* is the gradient norm clipping value.

## C Training Curves

In Figures 1, 2, and 3, we show the training curves of ROCES. The hard accuracy measures how a predicted expression is similar to a target class expression in the training set in terms of their string representation. The soft accuracy is the Jaccard index between the set of the predicted tokens and the set of the target tokens; it does not take the order of the tokens into account. The hard accuracy is hence the most suited metric to train ROCES. We refer to [Kouagou *et al.*, 2023] for more details about these metrics. We can observe a rapid increase in accuracy and decrease in loss in the early epochs and

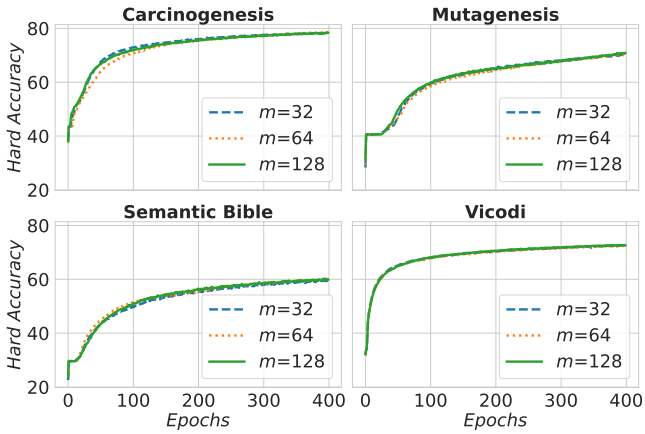



Figure 1: Hard accuracy curves of our proposed approach ROCES during training. $m$ is the number of inducing points in the Set Transformer model. The probability functions $p^+$, $p^-$ defined in Algorithm 1 in the main paper are used.

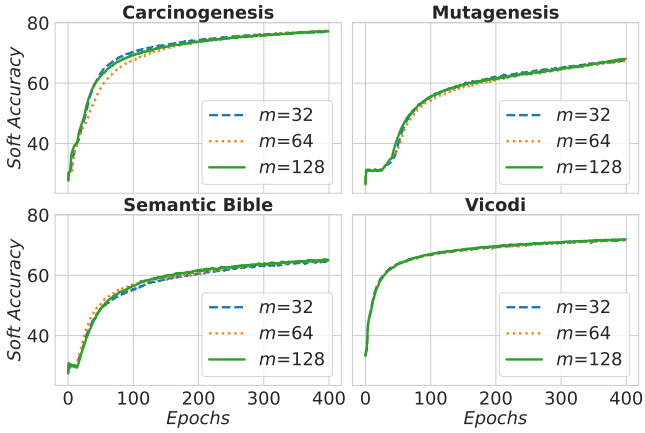



Figure 2: Soft accuracy curves during training

fast convergence on Carcinogenesis and Vicodi, which are the largest datasets. This observation aligns with the results presented in Table 2 (see main paper). This suggests that on large datasets, ROCES learns better mappings between sets of examples and class expressions that describe them.

# D Additional Results

## D.1 Distribution of the F-measure

In the experiment comparing ROCES against the best search-based approach EvoLearner, we let ROCES perform 50 trials and predict a solution for each trial, see Table 3 in the main paper. We now plot the distribution of the quality of the computed solutions for different learning problems.

From Figures 4a and 4b, we can observe that most of the solutions computed by ROCES have an F-measure close to 100% (see the large area around 100). This is more noticeable when considering the best performance

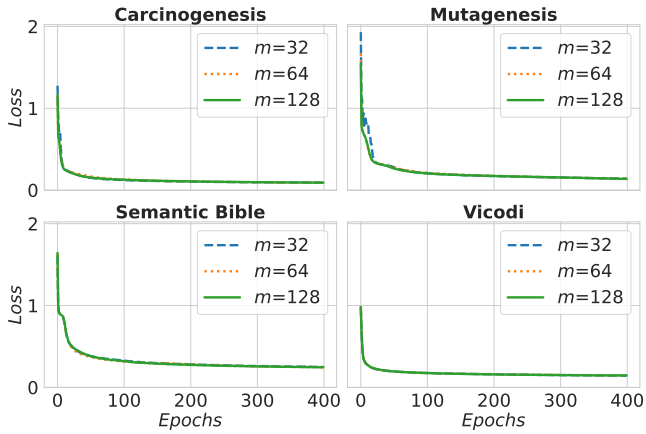



Figure 3: Loss (cross entropy) during training

across the 50 trials (see the distribution for the aggregation Max). Again, the highest performance is observed on the largest datasets Carcinogenesis and Vicodi, but also on Mutagenesis (when ROCES is allowed multiple trials). On Semantic Bible, we observe lower $F_1$ scores especially on the aggregation Min (see the large area around 0 for the blue violin plot). Nonetheless, when ROCES is allowed multiple trials, its performance significantly improves on this dataset (see the Max aggregation).

## D.2 Example Predictions

The solutions computed by different approaches for the learning problem discussed in Section 4.4 of the main paper are as follows:

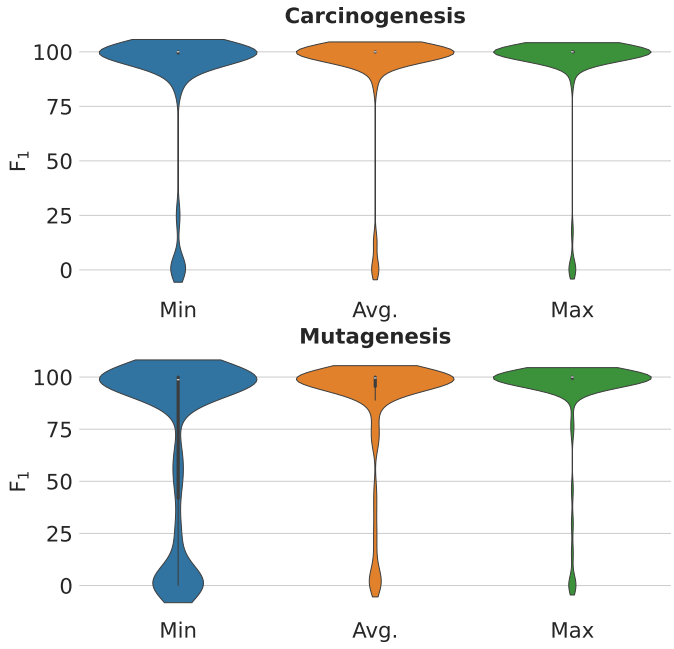
1. ROCES: $\texttt{Mountain} \sqcup (\texttt{GeopoliticalArea} \sqcap (\texttt{City} \sqcup (\exists\ \texttt{subregionOf}.\top)))$; $F_1$: 100%
2. ROCES$_U$: $\texttt{Mountain} \sqcup (\texttt{GeopoliticalArea} \sqcap (\texttt{City} \sqcup (\exists\ \texttt{subregionOf}.\top)))$; $F_1$: 100%
3. EvoLearner: $\texttt{City} \sqcup (\exists\ \texttt{location}.\top) \sqcup (\texttt{Mountain} \sqcap \texttt{GeographicArea}) \sqcup (\exists\ \texttt{subregionOf}.(\exists\ \texttt{subregionOf}.\texttt{GeographicArea}))$; $F_1$: 97.59%
4. CELOE: $\texttt{Mountain} \sqcup (\texttt{GeopoliticalArea} \sqcap \texttt{LandArea})$; $F_1$: 92.24%
5. CLIP: $\texttt{Mountain} \sqcup (\texttt{GeopoliticalArea} \sqcap (\texttt{City} \sqcup (\exists\ \texttt{subregionOf}.\top)))$; $F_1$: 100%
6. NCES2: $\texttt{Mountain} \sqcup (\texttt{GeopoliticalArea} \sqcap (\texttt{City} \sqcup (\exists\ \texttt{subregionOf}.\top)))$; $F_1$: 100%.

Although NCES2 and CLIP computed an exact solution with 100% $F_1$ score, they use all available examples for this purpose. This suggests that ROCES is competitive in performance even when it does not use all available examples.

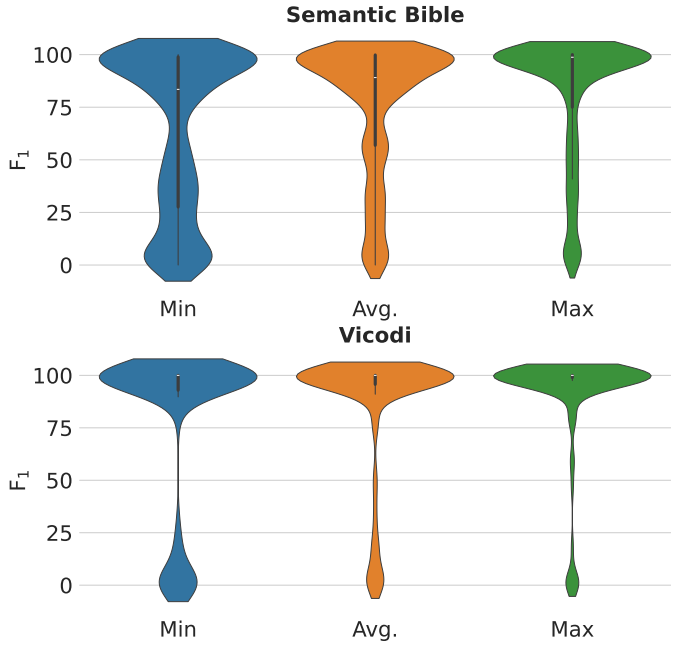
# References

[Kouagou *et al.*, 2023] N'Dah Jean Kouagou, Stefan Heindorf, Caglar Demir, and Axel-Cyrille Ngonga Ngomo. Neural class expression synthesis. In

*European Semantic Web Conference*, pages 209–226. Springer, 2023.




(a) Distribution of the F-measure on Carcinogenesis and Mutagenesis




(b) Distribution of the F-measure on Semantic Bible and Vicodi

Figure 4: Distribution of the F-measure of the solutions computed by ROCES for different aggregations. The aggregations (Min, Avg., Max) are computed across the 50 trials.