Reference-Free Evaluation of Taxonomies

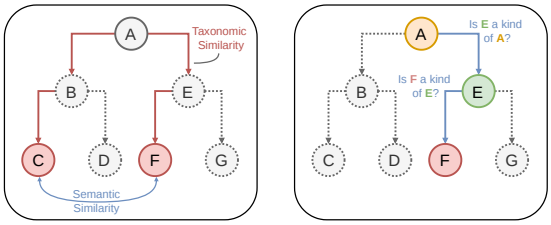
Pascal Wullschleger[*,†], **Majid Zarharan**[◇], **Donnacha Daly**[†]
Marc Pouly[†], **Jennifer Foster**[*]

[*] Hamilton Institute, Maynooth University, Ireland
[◇] School of Computing, Dublin City University, Ireland
[†] Lucerne School of Computer Science and IT, Switzerland

pascal.wullschleger@hslu.ch

Abstract

We introduce two reference-free metrics for quality evaluation of taxonomies in the absence of labels. The first metric evaluates robustness by calculating the correlation between semantic and taxonomic similarity, addressing error types not considered by existing metrics. The second uses Natural Language Inference to assess logical adequacy. Both metrics are tested on five taxonomies and are shown to correlate well with F1 against ground truth taxonomies. We further demonstrate that our metrics can predict downstream performance in hierarchical classification when used with label hierarchies.



(a) Robustness (b) Logical Adequacy

Figure 1: Core ideas of our reference-free measures. We correlate semantic and taxonomy similarity to measure robustness and check parent-child edges to assess logical adequacy of the taxonomy.

1 Introduction

Taxonomies are used to classify items, ideas or organisms based on shared characteristics. They help organize information, making it easier to find, understand and manage. Hierarchical classification, for example, enables systems to progressively categorize items from broad groups to specific subcategories, thereby simplifying complex classification tasks by leveraging the taxonomy of labels (Zangari et al., 2024). Likewise, taxonomies can be used to enrich features for machine learning (Choi et al., 2017; Ma et al., 2018; Wullschleger et al., 2022) or improve zero-shot classification (Geng et al., 2021; Kulmanov et al., 2019; Chen et al., 2020).

Taxonomies have traditionally been created manually, and, more recently, partially or fully automatically. Such automated taxonomy extension or generation methods address the challenges posed by ever-increasing volumes of unstructured data. By leveraging machine learning, systems can identify emergent categories, detect previously unrecognized relationships, and update or create taxonomic structures in real time with minimal human effort (Velardi et al., 2013; Jurgens and Pilehvar, 2016; Bordea et al., 2016; Zhang et al., 2021).

Taxonomy completion methods are usually evaluated by comparing the resulting taxonomies against ground truth taxonomies (Manzoor et al., 2020; Wang et al., 2022; Liu et al., 2021; Zhang et al., 2021; Xu et al., 2023; Shen et al., 2020; Shi et al., 2024; Zeng et al., 2025). In practice, however, there is no ground truth taxonomy to compare against when generating or completing taxonomies, because, if the ground truth taxonomy was available, there would be no need to expand or generate it. A potential solution is to use reference-free evaluation. We propose two reference-free measures, each targeting a particular aspect of taxonomic quality.

We build on established taxonomy quality criteria, particularly robustness (Unterkalmsteiner and Abdeen, 2023), but find existing automated measures do not account for non-leaf mutations, overlook error severity or heavily rely on priors within the backbone model's training data. In addition, criteria aimed at human-curated taxonomies overlook logical adequacy, such as whether a child concept is truly a subtype of its parent, presumably because people rarely make such errors. Consequently, we suggest two properties for evaluating taxonomy quality in a reference-free setting: the correlation of semantic and taxonomic similarity (*robustness* – see Fig. 1a) and the recognition of `is-a` parent-

child edges (*logical adequacy* – see Fig. 1b).

We evaluate our metrics on five taxonomies across various domains. By degrading ground truth taxonomies, we demonstrate that our metrics align with the level of deterioration, outperforming existing metrics. Similarly, we demonstrate that they outperform existing metrics in an extrinsic evaluation of taxonomy quality using the downstream application of hierarchical classification. Our source code is available on GitHub[1].

2 Background

2.1 Preliminaries

Following Zeng et al. (2021), a taxonomy $\mathcal{T} = (\mathcal{E}, \mathcal{V})$ is a directed acyclic graph with edges $\langle c_i, c_{i+1} \rangle \in \mathcal{E}$ pointing from a parent concept $c_i \in \mathcal{V}$ to a child concept $c_{i+1} \in \mathcal{V}$. Edges represent hypernym-hyponym pairs, where the child concept is the least detailed but different specialization of the parent concept. A placement of a concept is called a triplet (c_i, c_{i+1}, c_{i+2}), where c_{i+1} is the query concept that is placed as a child of c_i and parent of c_{i+2}. Following Manzoor et al. (2020), we add a pseudo-leaf and a pseudo-root to $\mathcal{V}$ to accommodate concepts without parents or children. If c_{i+1} is a leaf, c_{i+2} is the pseudo-leaf and if instead c_{i+1} is the root, then c_i is the pseudo-root. To compare a taxonomy to a ground truth, precision/recall/F1 are calculated over triplets.

As a measure of taxonomic similarity, we use the well-established Wu & Palmer Similarity (WPS) (Wu and Palmer, 1994). Let $p(c_k) = \langle c_0, ..., c_k \rangle$ denote the path from the root concept c_0 to a target concept c_k. Furthermore, let $\texttt{lca}(p(c_a), p(c_b))$ denote the depth of the least common ancestor of the paths $p(c_a)$ and $p(c_b)$. Then the WPS (Eq. 1) is the similarity between concepts c_a and c_b in the range $(0, 1]$, with 1 meaning that they are the same node.

$$W_{c_a c_b} = \frac{2 \cdot \texttt{lca}(p(c_a), p(c_b))}{|p(c_a)| + |p(c_b)|} \quad (1)$$

2.2 Taxonomy Quality Attributes

Unterkalmsteiner and Abdeen (2023) consolidate various names for taxonomy quality attributes found in prior work and identify a measurable minimal, well-defined set — comprehensiveness, robustness, conciseness, extensibility, explanatory, mutual exclusiveness, reliability (criteria definitions in App. A). We focus on *robustness*, which represents how well a taxonomy can tell things apart, meaning how clearly the concepts in a taxonomy represent different ideas and how closely related sibling concepts are. Robustness is independent of the application of the taxonomy and depends only on the concept space, since it describes an intrinsic quality. This makes it suitable for automated evaluation in a reference-free setting.

We also observe that prior work does not account for the need to verify `is-a` relationships. For instance, a parent *beverage* and a child *bread* would constitute an invalid relationship, since bread is not actually a type of beverage. While we assume such erroneous edges to be rare in human-created taxonomies, this does not necessarily hold for automatically generated taxonomies. We refer to this property as *logical adequacy*.

3 Methodology

This section outlines our methods for evaluating taxonomy robustness and logical adequacy. We also apply these and prior metrics to a toy example.

3.1 Robustness

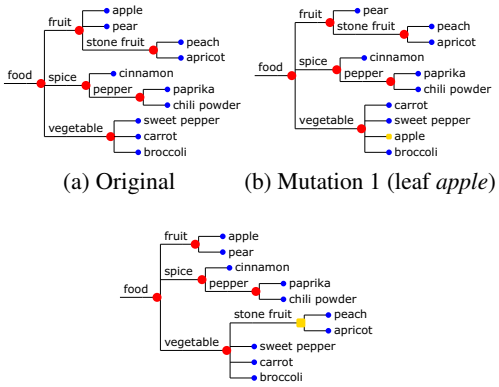
Fig. 2 shows two basic mutations of a toy taxonomy to which quality measures should be sensitive:

1. **Misclassified leaf concepts:** Move a leaf by randomly choosing a new parent (Fig. 2b).
2. **Misclassified non-leaf concepts:** Move a non-leaf and its descendants by randomly choosing a new parent. (Fig. 2c).

Unterkalmsteiner and Abdeen (2023) propose Semantic Proximity (SP) as a robustness metric that measures the rate of intruders in groups of siblings. They check if the minimum inter-group similarity is larger than the minimum extra-group similarity and average over all groups (see App. E). The SP metric disregards the hierarchical nature of taxonomies. It essentially treats groups of leaves as clusters, but does not consider that they could be nested in a hierarchical structure. We expect SP to be insensitive to misclassifications of non-leaf nodes (Fig. 2c) since it only looks at groups of leaf concepts to calculate a score.

To mitigate this, we propose *Concept Similarity Correlation (CSC)*, which is based on the assumption that the semantic distance between concepts and their taxonomic distance should correlate. We measure the WPS, $W_{c_i c_j}$, between concepts c_i

[1] https://github.com/wullli/reference-free-taxonomy-eval



(a) Original (b) Mutation 1 (leaf *apple*)

(c) Mutation 2 (non-leaf *stone fruit*)

Figure 2: Toy taxonomy to illustrate different mutations that a robustness metric should be sensitive to. Leaf nodes are marked in blue, mutated nodes in yellow.

and c_j using the paths from the root to the concepts (Eq. 1). Next, for each concept in the taxonomy, we define its semantic representation as the embedding of the concept description. Our CSC score is the Kendall rank correlation $\tau(.)$ of the semantic (cosine) similarity $S_{c_i c_j}$ between concept representations and the taxonomic similarity $W_{c_i c_j}$ (Eq. 2). Thus, we assume that the relationship between taxonomic and semantic similarity is monotonically increasing in a sensible taxonomy, i.e. if semantic similarity increases, so does taxonomic similarity. We calculate Kendall's τ using all pairs of concepts in the taxonomy (exhaustively).

$$\text{CSC} := \tau\left(S_{c_i c_j}, W_{c_i c_j}\right) \tag{2}$$

3.2 Logical Adequacy

Since robustness does not ensure valid parent-child edges, we propose a logical adequacy measure. Let us assume an external object that we want to classify using a taxonomy. Then, a classification process is a walk on the taxonomy graph represented as a tuple of edges $C = (\langle c_0, c_1\rangle, \ldots, \langle c_{k-1}, c_k\rangle)$ that connect the root concept c_0 to any concept c_k under which the object is classified, with $c_{i+1} = \texttt{child}(c_i)$ and $c_i, c_{i+1} \in \mathcal{V}$ for $i = 0, \ldots, k-1$.

We want the possible classifications (walks) to be probable (according to some model) for a high quality taxonomy, i.e. the parent-child edges of the walks need to be logically adequate. For analogy, in the context of language modelling, the objective is to generate probable or likely texts, implying that a probable text is considered a good one. Similarly, in our approach, parent-child relationships must be probable for a classification to be considered good. We do not simply score each edge and average, since we want mistakes at higher levels to impact the score more than e.g., misclassified leaves deep in the taxonomy, because they affect a larger subgraph. Thus, inner concept scores are deliberately used multiple times in the computation.

If A denotes the event of adequacy, we define the normalized probability of the classification C being adequate as the geometric mean of the edge probabilities (Eq. 3)

$$\tilde{P}(A|C) := \left[\prod_{i=0}^{k-1} P(A|\langle c_i, c_{i+1}\rangle)\right]^{\frac{1}{k}} \tag{3}$$

We express the adequacy of the whole taxonomy, $\tilde{P}(A)$, as the normalised probability that a randomly sampled classification in our taxonomy is adequate, which is the union probability across classifications. This results in the mean over probabilities of adequacy given the classification (Eq. 4). Refer to App. C for further detail.

$$\tilde{P}(A) := \sum_C \tilde{P}(A|C) \times P(C) = \frac{1}{|\mathcal{V}|} \sum_C \tilde{P}(A|C) \tag{4}$$

Approximating Edge Probabilities We estimate $P(A|\langle c_i, c_{i+1}\rangle)$, the probability of an edge being adequate, using Natural Language Inference (NLI). We rely on a flavour of NLI that approximates a probability distribution over the classes *contradicts*, *neutral* and *entails*, regarding whether a premise entails a hypothesis. For example, consider the edge (*appetizer*, *antipasto*) in a taxonomy of food items. We define the premise as the description of the child concept, e.g., *"antipasto is a course of appetizers in an Italian meal"* and the hypothesis as a string characterising the is-a relationship between the child and parent nodes in an edge, e.g. *"antipasto is a type of appetizer"*. The expression $s(\langle c_i, c_{i+1}\rangle)$ denotes the string of the premise and hypothesis resulting from the application of the template in Eq. 5 to the edge, where $D(c_i)$ denotes the description of the concept c_i, $N(c_i)$ the lemma of the concept name, and $A(c_i)$ the appropriate article (e.g., *a*, *an*, *the*). We abbreviate metrics based on this NLI-approximation as NLIV (NLI Verification).

$$s(\langle c_i, c_{i+1}\rangle) = \tag{5}$$
$$''D(c_{i+1}).\, A(c_{i+1})\, N(c_{i+1}) \text{ is a type of } N(c_i)''$$

There are two possibilities for estimating the probability that an edge is adequate:

1. **Strong:** the premise must *entail* the hypothesis (*NLIV-S*) as shown in Eq. 6.

$$P(A \mid \langle c_i, c_{i+1} \rangle) \approx \\ NLI(Y = \texttt{entails} \mid s(\langle c_i, c_{i+1} \rangle)) \quad (6)$$

2. **Weak:** the premise must *not contradict* the hypothesis (*NLIV-W*) as shown in Eq. 7.

$$P(A \mid \langle c_i, c_{i+1} \rangle) \approx \\ 1 - NLI(Y = \texttt{contradicts} \mid s(\langle c_i, c_{i+1} \rangle)) \quad (7)$$

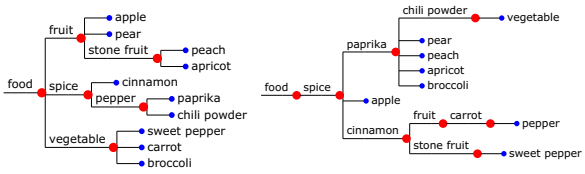
Hearst Patterns Instead of only relying on a single prompt template for NLI-approximation, we use the mean estimate over instances of Hearst patterns (Hearst, 1992), which were originally designed to extract hypernym-hyponym edges from natural language texts using regular expressions. The first pattern is presented below (complete list in App. D):

$$s(\langle c_i, c_{i+1} \rangle) \in S,\ S = \{ \\ ''D(c_{i+1}).\ A(c_{i+1})\ N(c_{i+1}) \text{ is a type of } N(c_i)'', \\ ''D(c_{i+1}).\ A(c_{i+1})\ N(c_{i+1}) \text{ is an example of } N(c_i)'', \\ ''D(c_{i+1}).\ A(c_{i+1})\ N(c_{i+1}) \text{ is } A(c_i)\ N(c_i)'', \\ ''D(c_{i+1}).\ A(c_{i+1})\ N(c_{i+1}) \text{ is a kind of } N(c_i)''\}$$

Parallels to Previous Work Langlais and Gao (2023) propose rating parent-child edges by generating parent concepts for a given child using masked language model prompts and assessing whether the true parent appears in the returned set. While they average over all parent-child pairs, we work with complete classifications. We compare to their metric, RaTE, in our experiments.

3.3 Toy Example

The first two rows of Tab. 1 show a comparison of all reference-free metrics on the toy example from Fig. 2. We can see that SP is only sensitive to Mutation 1, where a leaf concept is moved, and not to Mutation 2, where a non-leaf is moved. Our proposed metrics, CSC and NLIV, on the other hand, are sensitive to both mutations.



(a) Best in sample (b) Worst in sample

Figure 3: The best and worst toy taxonomies according to the product of CSC and NLIV-S in a 10k sample of the toy example space.

Mutation	SP [0, 1]	CSC [−1, +1]	NLIV-S [0, 1]	NLIV-W [0, 1]	RaTE [0, 1]	CSC×NLIV-S [0, 1]
Leaf (1)	0.4861	0.3745	0.5303	0.9854	0.8000	0.1986
Non-Leaf (2)	0.7805	0.3272	0.5087	0.9873	0.8000	0.1665
Worst in sample	0.0625	-0.2126	0.3440	0.9409	0.2667	-0.0731
Best in sample	0.7805	0.4467	0.5604	0.9909	0.8000	0.2504
Original	0.7805	0.4467	0.5604	0.9909	0.8000	0.2504

Table 1: Results on the toy example. The value range of metric scores is indicated in brackets.

For an intuition of what is considered bad by the proposed metrics, Fig. 3 shows the best and worst taxonomies found by sampling 10k taxonomies in the concept space of the toy example. We do not show exhaustive search since even if we only consider trees (as opposed to DAGs), Cayley's formula states that the number of possible taxonomies for the toy example is already $n^{n-2} \approx 7.21 \times 10^{16}$ (Shor, 1995). The metric that we use for determining what is best and what is worst is the product of NLIV-S and CSC. The middle two rows of Tab. 1 show the scores for the best and worst taxonomy according to all metrics. The taxonomy resulting in the best scores is the original toy taxonomy while the worst scores belong to a taxonomy where related concepts are far apart and hypernym-hyponym edges are invalid.

4 Metric Evaluation

4.1 Intrinsic Evaluation

Data Following Xu et al. (2023) and Wang et al. (2022), we evaluate on SemEval-Food (SF), SemEval-Verb (SV), and MeSH (ME) datasets. SemEval-Food, the largest taxonomy from SemEval-2016 Task 13, was used to evaluate taxonomy extraction methods (Bordea et al., 2016). SemEval-Verb, based on WordNet 3.0 (Fellbaum, 2019), was featured in SemEval-2016 Task 14 for evaluating taxonomy enrichment approaches (Jurgens and Pilehvar, 2016). MeSH is a hierarchical vocabulary of medical terms (Lipscomb, 2000).

In the food industry, taxonomies play a critical role in the development of new recipes and their adaptation to changing culinary trends, dietary requirements, and sustainability objectives. We therefore experiment with two further taxonomies in this domain. We extract a taxonomy from Wikidata[2] (WT) by using `Food (Q2095)` as the root and retrieving all descendants (see App. F.3), and we use a proprietary taxonomy from a major food market

[2] https://www.wikidata.org/

Dataset	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	D	$\lvert\mathbf{L}\rvert$	$\frac{\lvert\mathbf{L}\rvert}{\lvert\mathcal{V}\rvert}$	B
MeSH (ME)	9,710	10,496	11	5,502	0.57	3.88
CookBook (CB)	1,985	1,984	4	1,795	0.90	10.44
Wikitax (WT)	941	973	7	754	0.80	5.20
SemEval-Food (SF)	1,486	1,576	9	1,184	0.80	5.08
SemEval-Verb (SV)	13,936	13,407	13	10,360	0.74	4.12

Table 2: Benchmark taxonomy statistics. $|\mathcal{V}|$, $|\mathcal{E}|$, D, $|L|$, $\frac{|L|}{|\mathcal{V}|}$, B represent the concept number, edge number, depth, leaf number, leaf ratio and branching factor

chain employed in recipe development, which we call the *CookBook* (CB) taxonomy.[3] For details on the datasets, see Tab. 2.

Pre-Trained Models As an NLI-model, we use BART-L (Lewis et al., 2020), which used the same pre-training data as Liu and Lapata (2019), including BooksCorpus (Zhu et al., 2015), English Wikipedia (text only) and various news-article datasets, and was then post-trained for zero-shot NLI on the MultiNLI dataset (Williams et al., 2018). NLI models are only used in zero-shot mode in our metrics, thus we validate them only in this mode. Zero-shot is also the most challenging setting and will therefore provide a lower bound for performance. For semantic similarity, we use MiniLM-L6, which is a sentence-transformer that generates embeddings of the mean over non-padding token representations (Reimers and Gurevych, 2019). See App. F.2 for the full model identifiers.

Metric Evaluation To evaluate the proposed metrics, we conduct an empirical validation using our benchmark taxonomies, which are assumed to be of high quality and to reflect human judgment. We degrade all taxonomies by randomly relocating subgraphs of the original taxonomies according to the mutations described in Sec. 3.1. To this end, we mutate between 2% and 32% of the taxonomy nodes and calculate F1 (against the original taxonomy), NLIV and CSC (our proposed metrics) as well as RaTE and SP (previous metrics) every $2^x\%$ mutations with $x = 1, \ldots, 5$. Depending on the taxonomy, we sample between 50-100 different degradations, resulting in 250-500 degraded taxonomies. We then calculate the correlation between F1 and CSC/NLIV-S/NLIV-W/RaTE/SP. Note that mutations are sequential and a subsequent mutation can affect a previously mutated subgraph.

We present two mutation settings. In the *non-leaf* setting, we only move sub-graphs of inner concepts (not leaves). In the *all* setting, we allow any concept's sub-graph to be moved. Our NLIV metrics assume that mistakes higher in the taxonomy are more severe than mistakes at lower levels, since they affect more descendants. Thus, an unweighted F1-score is biased against our metric. For completeness, we calculate F1 scores weighted by the number of descendants of a node in addition to an unweighted score.

Results Fig. 4 shows Kendall's τ between the F1 score against the ground truth taxonomy and the reference-free metric scores. Correlations are undefined (NA) when the metric predicts the same score

[3]Both taxonomies are available with our source code at https://github.com/wullli/reference-free-taxonomy-eval.

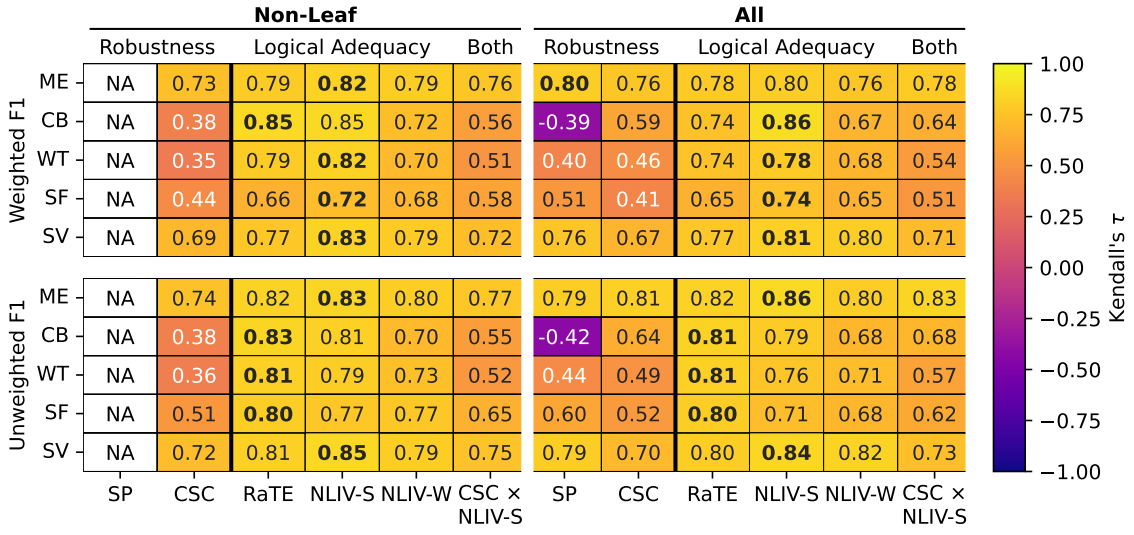


Figure 4: Correlations between F1 and reference-free metrics, including the correlations when only mutating non-leaves. All correlations are statistically significant with $\alpha = 0.001$. We calculated F1 scores with and without class weights, assigning weights based on the number of descendants of a concept.

for all mutated taxonomies. For instance, when mutating only non-leaf concepts, SP remains constant and the correlations are thus undefined (NA) for all datasets. Our metrics, CSC and NLIV-S, are strongly associated with F1 across datasets. For leaf nodes, SP performs effectively except on CookBook, likely due to its higher branching factor (see Tab. 2). SP demands that intra-group similarity be below extra-group similarity per leaf, but larger leaf groups increase intruder risk due to lower cohesion. Although CSC×NLIV-S has a weaker correlation than NLIV-S alone, it helps avoid flat taxonomies. Our degraded taxonomies are unlikely to be flat, given an expected branching factor of 2 for non-leaves in the limit (Zhang and Zhang, 2015). NLIV-W has a weaker correlation with F1, likely because the model does not always perceive invalid edges as contradictory. RaTE shows consistently high correlations, but, as expected, falls short of NLIV-S in the weighted F1. RaTE also calculates scores using only a subset of the taxonomy, since for most concepts the true parent is not actually in the top-k and the score is simply zero over this subset ($\sim 50\% - 90\%$, see Fig. 12 in App. G).

4.2 Extrinsic Evaluation

We next evaluate how our metrics correlate with performance on a downstream task. We assume that the quality of a taxonomy can be characterised by how well external objects can be classified into it, i.e. how well it performs in a hierarchical classification setting. For example, in hierarchical text classification, a text must be categorised into predefined hierarchically organised classes, with general classes at the higher levels and detailed classes at lower levels (see Tab. 3). This task is well-suited for extrinsic evaluation because a correct class hierarchy (taxonomy) is crucial in preventing error propagation. That is, since high-level predictions influence low-level ones, a low-level target misplaced in the hierarchy is harder to predict.

Data We experiment with three hierarchical classification datasets, each accompanied by a label taxonomy. For data statistics, refer to Tab. 4. Web Of Science (WOS) includes abstracts of scientific articles that must be classified into their respective topic domains (level 1) and areas (level 2) (Kowsari et al., 2017). The Multilabeled News Dataset (MN-DS) is a collection of news articles gathered over more than a year from an academic news search engine (Petukhova and Fachada, 2023). The DBPedia (DBP) dataset[4] is based on the DBPedia project (Auer et al., 2007), which aims to extract structured information from Wikipedia. The dataset uses the DBPedia ontology to provide hierarchical labels for a large collection of Wikipedia articles.

Article	Taxonomic Labels		
	Lvl. 1	Lvl. 2	Lvl. 3
William Alexander Massey (October 7, 1856 – March 5, 1914) . . .	Agent	Politician	Senator
Lions is the sixth studio album by American rock band The . . .	Work	MusicalWork	Album
Pirqa (Aymara and Quechua for wall, hispanicized spelling . . .	Place	NaturalPlace	Mountain

Table 3: Examples from the DBP dataset.

Dataset	# Samples	Lvl. 1	Lvl. 2	Lvl. 3
Web Of Science (WOS)	46,985	7	134	-
MN-DS News (MN-DS)	10,917	17	109	-
DBPedia (DBP)	342,782	9	70	219

Table 4: Hierarchical classification dataset statistics The levels (Lvl. 1-3) indicate the number of labels per taxonomic depth, where 0 indicates the root.

We divide WOS and MN-DS into 80% training and 20% test, while keeping the published data split for DBPedia (∼18% test, ∼10% validation). We do not use the validation set.

Classification Approach We use logistic regression classifiers on top of sentence-transformer embeddings (`MiniLM-L6`) of the article texts. We perform standard top-down prediction by fitting one classifier per parent node (Zangari et al., 2024). At each level, one of the children of the previously selected node is predicted in a multi-class setting. This results in a number of classifiers equal to the number of inner nodes plus the root.

Setup Following the intrinsic evaluation experimental setup, we degrade versions of all label taxonomies (WOS, DBP, MN-DS) by randomly relocating sub-graphs of the original label taxonomies. We again mutate between 2% and 32% of the taxonomy nodes and calculate the downstream F1 score by fitting the hierarchical classification outlined above using the degraded label taxonomy. As before, reference-free metric scores are calculated every 2^x mutations with $x = 1, \ldots, 5$. We sample 100 different degraded taxonomies for each dataset. The correlation is then calculated between the macro F1 score on the lowest-level taxonomic

[4] https://huggingface.co/datasets/DeveloperOats/DBPedia_Classes

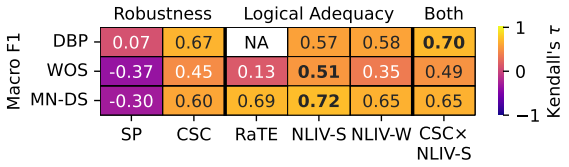


Figure 5: Correlations between macro F1 on the downstream task and our reference-free metrics. All correlations are statistically significant with $\alpha = 0.001$.

labels and the reference-free metrics. The assumption is therefore that the classifier performance is associated with the quality of the label taxonomy. We use the macro F1, since the datasets are not balanced and we want to give equal importance to all classes.

Results Fig. 5 shows the correlations between macro F1 on the downstream task and reference-free metrics. All versions of our metrics show substantial associations with the downstream performance, whereas RaTE and SP do not, or do so only partially. On DBP, RaTE never predicts the true parent of a node in the top-k, resulting in a constant score and therefore an undefined correlation. This is one of the reasons Langlais and Gao (2023) perform domain-specific fine-tuning of masked language models. SP is even negatively associated with downstream performance on WOS and MN-DS, likely because of the high branching factor of the ground truth label taxonomy, which makes intruders in sibling groups very likely.

For an intuition on how the metrics behave, we show scatter plots and a fitted exponential on the space of the collected metric scores of taxonomies mutated during the experiments (Fig. 14 in App. G.1). CSC and NLIV can, to an extent, predict downstream F1, as indicated by the R^2. While RaTE shows a positive tendency, there is higher variance and it does not work for DBP without tuning as shown in Fig. 5. The scatter plot for SP clearly shows its negative association with F1.

4.3 Extended Analysis

We carry out further analysis of underlying assumptions and modelling decisions by subjecting our metrics to variations in input and base models.

Realistic Errors In the experiments in Section 4.1, we degrade taxonomies by moving either *non-leaves* or *all* concepts. This has the advantage of making no assumptions about the error modes of potential generation or completion algorithms.

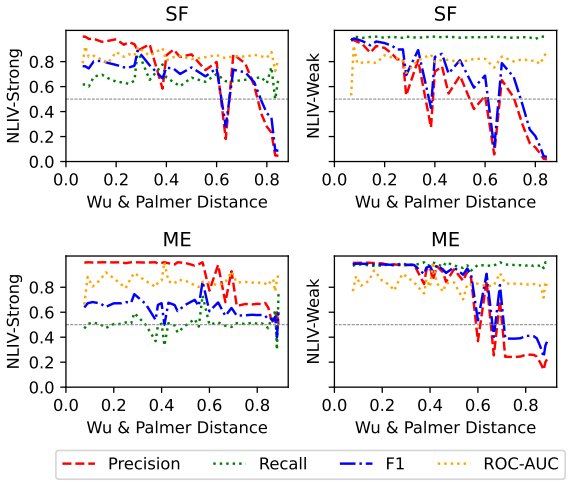


Figure 6: Binary classification performance with zero-shot NLI on the SF and ME taxonomies. The x-axis is the WPS between the two concepts in the ground truth taxonomy. The gray dashed line is the random baseline.

However, since errors in taxonomy generation or completion methods often place nodes under similar but wrong hypernyms, we conduct an additional intrinsic evaluation, where a concept's sub-graph can only be moved under one of the top 100 closest concepts according to WPS[5]. The results are shown in Fig. 7. The correlations for the NLIV metrics remain high but are lower for CSC, which is expected. CSC, by definition, focusses on the global shape of the taxonomy, and moving concepts locally will only have a minor effect on it.

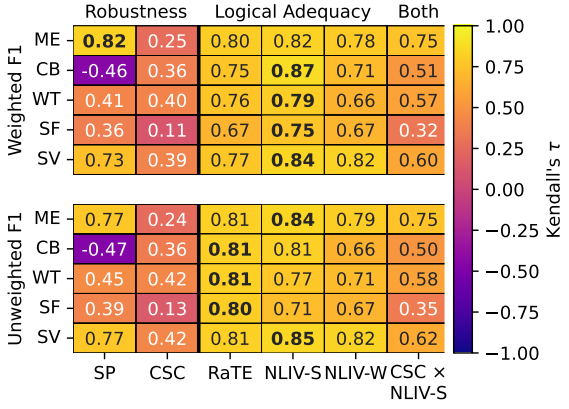


Figure 7: Correlations between F1 and reference-free metrics when moving a node in its vicinity (weighted by WPS). All correlations are significant with $\alpha = 0.001$.

Approximating Parent-Child Likelihoods Since our logical adequacy metric depends on NLI to estimate the probability that a parent-child edge

[5]WPS is also used as sampling weight.

is adequate, it would be undermined by low NLI accuracy. We therefore define a binary classification task to evaluate the NLI model's suitability. We use existing edges as positive instances and an equal amount of random, unrelated concept pairs as negatives. We consider a strong (NLIV-S, Eq. 6) and weak (NLIV-W, Eq. 7) version. Fig. 6 shows the NLI binary classification performance on SemEval-Food (SF) and MeSH (ME). See Tab. 7 in App. G for the full set of scores[6]. As expected, we observe higher precision compared to recall with NLIV-S, while the opposite is true for NLIV-W. The ROC-AUC hovering around 0.8 for both datasets and metrics indicates that the NLI models are indeed quite discriminative and thus applicable to our metrics. Scores change as the two concepts of a pair become more unrelated, implying that the decision boundary depends on the taxonomic distance of a pair.

Sensitivity to Input Perturbations To test how stable our metrics are to different, semantically identical taxonomies, we reuse our intrinsic evaluation setup, but for each mutated taxonomy, we create a corresponding taxonomy that has between 0% and 100% of its concepts replaced by synonyms. We map concept names to WordNet (Fellbaum, 2019) where possible and sample a synonym from the synset of the most common meaning. We then calculate scores for both taxonomies and compare the rank correlations using Kendall's τ. If our metrics are sufficiently robust to the induced perturbations, we expect large τ's between mutated taxonomies and their perturbed versions. The results are shown in Fig. 8. While CSC is robust with high Kendall's τ between the two distributions, the NLIV metrics are more sensitive to changes in inputs, which aligns with the literature on NLI models (Arakelyan et al., 2024). However, NLIV metrics still show relatively high association between the original and perturbed taxonomies when compared to RaTE (see plot for RaTE and SP in Fig. 11, App. G). The stability of NLIV also seems to be dataset-dependent, as indicated by the differences in τ for the same metric.

Sensitivity to Model Changes To test how robust our methodology is to changes in model, we evaluated four different sentence transformers (MiniLM-L6, BGE-M3, ML-E5-L, and Para-MPNet) and NLI-models (RoBERTa-L, DeBERTa-L, BART-L, DeBERTa-ML) against the same taxonomy mutations and compared the ranking of the taxonomies using Kendall's τ. The results are shown in Fig. 9. Scores cannot be directly compared between models due to the shifting distributions, but the consistently high τ suggests that the same taxonomies can be compared using different models and the ranking will stay fairly consistent.

[6]We also tried using concept names instead of descriptions and found that performance is almost entirely conserved.

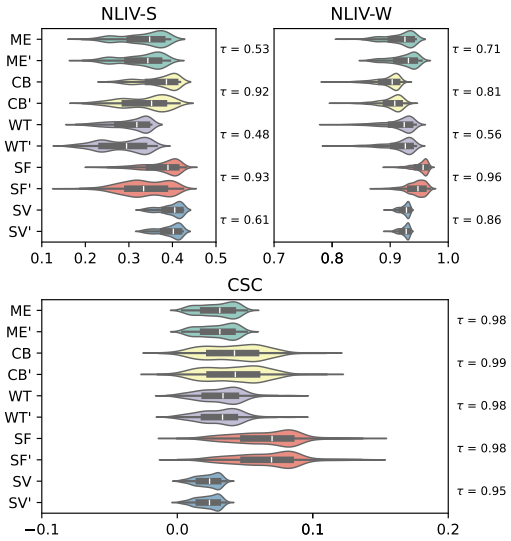


Figure 8: Violin plots showing the distribution of our scores when replacing concept names with synonyms using WordNet. The right hand side shows the Kendall's τ between the two distributions as a measure of their monotonic association. The distributions of perturbed taxonomies are indicated with primes (e.g., CB′).

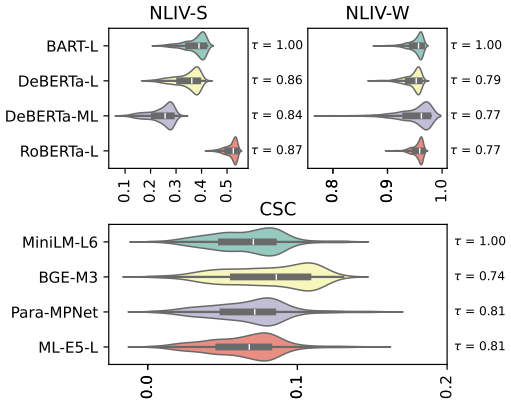


Figure 9: Violin plots showing the distribution of our scores when changing backbone models on SemEval-Food. The righthand side shows the Kendall's τ between the model score distribution and the models used in our original experiments (BART-L and MiniLM-L6).

5 Conclusion

We propose two novel metrics for assessing taxonomy quality, CSC and NLIV, that offer greater resilience across domains and mutations than existing ones without requiring fine-tuning. We show that they are highly associated with taxonomy quality as measured by comparison to gold taxonomies, and are predictive of hierarchical classification performance when applied to label taxonomies. We envisage using these metrics primarily to compare alternative taxonomies, but they could also be used to pinpoint individual errors by, for example, using classification probabilities predicted by NLIV to identify very unlikely classification walks, or examining concept pairs where semantic similarity and taxonomic similarity show the highest disagreement. Further research is needed to explore these possibilities, as well as to investigate how these metrics could be leveraged as an optimization objective or how they relate to human intuition about taxonomy quality.

6 Limitations

- While our proposed metrics are designed to be domain independent, the pre-trained models we use are not. As a result, the computed scores may reflect biases inherent in the models used for semantic similarity and NLI.
- Our assessment of logical adequacy depends on NLI. Before applying our method, we recommend evaluating the model performance in the taxonomy domain through experiments similar to those in Section 4.3.
- While we observe in experiments that the metrics can differentiate taxonomy quality within the same concept space, we do not expect our metrics to be directly comparable between different concept spaces due to the varying performances of baseline models in these spaces and the potentially differing goals and properties of their taxonomies.

7 Acknowledgements

We would like to express our sincere appreciation to Betty Bossi[7] for their support of this research project and for providing us with their taxonomy used for recipe development. This research is supported through computing resources by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2) and is co-funded under the European Regional Development Fund (ERDF). The authors thank the reviewers for their insightful and helpful comments. One author was supported in part by the Swiss National Science Foundation (SNSF) under grant 20HW-1_228541.

[7] https://www.bettybossi.ch/

References

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444, St. Julian's, Malta. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jiaoyan Chen, Freddy Lécué, Yuxia Geng, Jeff Z. Pan, and Huajun Chen. 2020. Ontology-guided Semantic Composition for Zero-shot Learning. In *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning*, pages 850–854.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. Gram: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 787–795, New York, NY, USA. Association for Computing Machinery.

Christiane Fellbaum, editor. 2019. *Wordnet*. Language, Speech and Communication. Bradford Books, Cambridge, MA.

Yuxia Geng, Jiaoyan Chen, Zhuo Chen, Jeff Z. Pan, Zhiquan Ye, Zonggang Yuan, Yantao Jia, and Huajun Chen. 2021. Ontozsl: Ontology-enhanced zero-shot learning. In *Proceedings of the Web Conference 2021*,

WWW '21, page 3325–3336, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

David Jurgens and Mohammad Taher Pilehvar. 2016. SemEval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1092–1102, San Diego, California. Association for Computational Linguistics.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 364–371.

Maxat Kulmanov, Wang Liu-Wei, Yuan Yan, and Robert Hoehndorf. 2019. El embeddings: Geometric construction of models for the description logic el++. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6103–6109. International Joint Conferences on Artificial Intelligence Organization.

Phillippe Langlais and Tianjian Lucas Gao. 2023. RaTE: a reproducible automatic taxonomy evaluation by filling the gap. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 173–182, Nancy, France. Association for Computational Linguistics.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Carolyn E. Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265–266. Historical Article.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zichen Liu, Hongyuan Xu, Yanlong Wen, Ning Jiang, HaiYing Wu, and Xiaojie Yuan. 2021. TEMP: Taxonomy expansion with dynamic margin loss through taxonomy-paths. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3854–3863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 743–752, New York, NY, USA. Association for Computing Machinery.

Emaad Manzoor, Rui Li, Dhananjay Shrouty, and Jure Leskovec. 2020. Expanding taxonomies with implicit edge semantics. In *Proceedings of The Web Conference 2020*, WWW '20, page 2044–2054, New York, NY, USA. Association for Computing Machinery.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Alina Petukhova and Nuno Fachada. 2023. Mn-ds: A multilabeled news dataset for news articles hierarchical classification. *Data*, 8(5).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020. Taxoexpan: Self-supervised taxonomy expansion with

position-enhanced graph neural network. In *Proceedings of The Web Conference 2020*, WWW '20, page 486–497, New York, NY, USA. Association for Computing Machinery.

Jingchuan Shi, Hang Dong, Jiaoyan Chen, Zhe Wu, and Ian Horrocks. 2024. Taxonomy completion via implicit concept insertion. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2159–2169, New York, NY, USA. Association for Computing Machinery.

Peter W Shor. 1995. A new proof of cayley's formula for counting labeled trees. *Journal of Combinatorial Theory, Series A*, 71(1):154–158.

Michael Unterkalmsteiner and Waleed Abdeen. 2023. A compendium and evaluation of taxonomy quality attributes. *Expert Systems*, 40(1):e13098.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. OntoLearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Suyuchen Wang, Ruihui Zhao, Yefeng Zheng, and Bang Liu. 2022. Qen: Applicable taxonomy completion via evaluating full taxonomic relations. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1008–1017, New York, NY, USA. Association for Computing Machinery.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 133–138, USA. Association for Computational Linguistics.

Pascal Wullschleger, Simone Lionetti, Donnacha Daly, Francesca Volpe, and Grégoire Caro. 2022. Autoregressive self-attention models for diagnosis prediction on electronic health records. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1950–1956.

Hongyuan Xu, Ciyi Liu, Yuhang Niu, Yunong Chen, Xiangrui Cai, Yanlong Wen, and Xiaojie Yuan. 2023. TacoPrompt: A collaborative multi-task prompt learning method for self-supervised taxonomy completion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15804–15817, Singapore. Association for Computational Linguistics.

Alessandro Zangari, Matteo Marcuzzo, Matteo Rizzo, Lorenzo Giudice, Andrea Albarelli, and Andrea Gasparetto. 2024. Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7).

Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Zhenyu Wu, Shangbin Feng, and Meng Jiang. 2025. CodeTaxo: Enhancing taxonomy expansion with limited examples via code language prompts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4131–4144, Vienna, Austria. Association for Computational Linguistics.

Qingkai Zeng, Jinfeng Lin, Wenhao Yu, Jane Cleland-Huang, and Meng Jiang. 2021. Enhancing taxonomy completion with concept generation via fusing relational representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 2104–2113, New York, NY, USA. Association for Computing Machinery.

Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. 2021. Taxonomy completion via triplet matching network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4662–4670.

Yazhe Zhang and Y. Z. Zhang. 2015. On the number of leaves in a random recursive tree. *Brazilian Journal of Probability and Statistics*, 29(4):897–908.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.

A Quality Criteria

The consolidated quality criteria of taxonomies according to Unterkalmsteiner and Abdeen (2023) are detailed below.

- **Comprehensiveness**: the capacity to categorize all objects within the taxonomy's target domain.
- **Robustness**: how well a taxonomy separates concepts, depending on orthogonality (concepts represent distinct ideas) and cohesiveness (sibling concepts are closely related).
- **Conciseness**: the capacity to classify objects using a minimal number of concepts; an external quality observed in practical application.
- **Extensibility**: the capability to accommodate structural modifications such as adding, altering, or removing roots or concepts.
- **Explanatory**: allows users to classify objects by their characteristics or infer characteristics from placement.
- **Mutual exclusiveness**: uniquely identifies an object, ensuring no object falls under different concepts within the same dimension; an external quality observed during application.
- **Reliability**: ensures consistent classification across different coders.

B Wu & Palmer Similarity

Fig. 10 illustrates how to calculate the WPS (Eq. 8) using an example.

$$\mathrm{W}_{c_a c_b} = \frac{2 \cdot \texttt{lca}(p(c_a), p(c_b))}{|p(c_a)| + |p(c_b)|} \tag{8}$$

C Logical Adequacy

The joint probability (Eq. 9) becomes the product of edge probabilities due to the assumption that edge probabilities are independent.

$$P(A|C) := \prod_{i=0}^{k-1} P(A|\langle c_i, c_{i+1}\rangle) \tag{9}$$

Our inspiration for normalising the probability by length (Eq. 10) using the geometric mean stems from the well known perplexity measure (Jelinek et al., 2005).

$$\tilde{P}(A|C) := \left[\prod_{i=0}^{k-1} P(A|\langle c_i, c_{i+1}\rangle)\right]^{\frac{1}{k}} \tag{10}$$

Since we have no notion of what the distribution of classifications looks like during the application of the taxonomy, we define it as uniform to avoid favouring any specific use case. This means $P(C)$ of Eq. 11 becomes $\frac{1}{|\mathcal{V}|}$ for all C.

$$\tilde{P}(A) := \sum_C \tilde{P}(A|C) \times P(C) = \frac{1}{|\mathcal{V}|} \sum_C \tilde{P}(A|C) \tag{11}$$

D Hearst Patterns

Here we show the full list of pattern instances used for the approximation of edge adequacy in the NLIV metrics.

- **Pattern (1):**

  $s(\langle c_i, c_{i+1}\rangle) \in S,\ S = \{$
  $''D(c_{i+1}).A(c_{i+1})\ N(c_{i+1})$ is a type of $N(c_i)''$,
  $''D(c_{i+1}).A(c_{i+1})\ N(c_{i+1})$ is an example of $N(c_i)''$,
  $''D(c_{i+1}).A(c_{i+1})\ N(c_{i+1})$ is $A(c_i)\ N(c_i)''$,
  $''D(c_{i+1}).A(c_{i+1})\ N(c_{i+1})$ is a kind of $N(c_i)''\}$

- **Pattern (2):**

  $s(\langle c_i, c_{i+1}\rangle) \in S,\ S = \{$
  $''D(c_{i+1}).A(c_i)\ N(c_i)$ such as $A(c_{i+1})\ N(c_{i+1})''$,
  $''D(c_{i+1})$.Such $N(c_i)_{\text{plural}}$ as $N(c_{i+1})''\}$

- **Pattern (3):**

  $s(\langle c_i, c_{i+1}\rangle) =$
  $''D(c_{i+1}).A(c_{i+1})\ N(c_{i+1})$ or other $N(c_i)''_{\text{plural}}$

- **Pattern (4):**

  $s(\langle c_i, c_{i+1}\rangle) =$
  $''D(c_{i+1}).A(c_{i+1})\ N(c_{i+1})$ and other $N(c_i)''_{\text{plural}}$

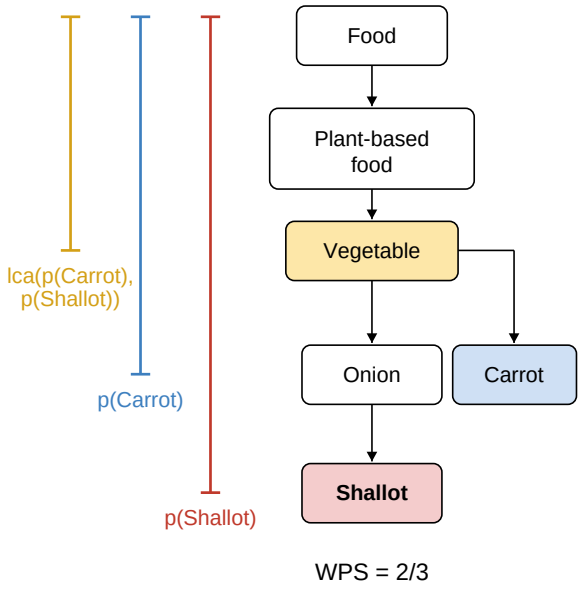


Figure 10: Example of WPS calculation.

- **Pattern (5):**

$$s(\langle c_i, c_{i+1}\rangle) = \\ ''D(c_{i+1}).N(c_i)_{\text{plural}},\ \text{including}\ N(c_{i+1})''$$

- **Pattern (6):**

$$s(\langle c_i, c_{i+1}\rangle) = \\ ''D(c_{i+1}).N(c_i)_{\text{plural}},\ \text{especially}\ N(c_{i+1})''$$

E Semantic Proximity

Eq. 12 shows the SP where $\mathcal{V}$ is the set of all concepts, $\mathcal{L}$ denotes the set of all leaf concepts in $\mathcal{V}$, $\mathcal{G}(l)$ is the set of sibling leaves of the leaf l, i and j represent leaves, and $\texttt{sim}(i, j)$ denotes the similarity measure between the two. The indicator function $\mathbb{I}(\cdot)$ evaluates to 1 if the expression inside is true and to 0 if it is false. This formulation, although different in notation, is equivalent to the original formulation in the work of Unterkalmsteiner and Abdeen (2023), but expressed in terms of sets.

$$SP = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}} 1 - \mathbb{I}\left(\min_{i,j\in\mathcal{G}(l), i\neq j} \texttt{sim}(i,j) > \min_{k\in\mathcal{V}\setminus\mathcal{G}(l)} \texttt{sim}(l,k)\right) \tag{12}$$

F Experiment Details

F.1 Degrading Taxonomies Randomly

We sample two unrelated concepts (neither are descendants or ancestors of each other) in the taxonomy and change the parent of the first concept to the second concept. In some taxonomies, concepts can have multiple parents. In such cases, we remove all current parents before adding the concept at the new parent. Tab. 5 shows the number of sampled degraded taxonomies per taxonomy. For the experiments in the extended analysis, Tab. 5 holds as well. The extrinsic evaluation features 100 degradations for each dataset; therefore, there are a total of 500 samples (5 for each degradation).

F.2 Model Identifiers

Tab. 6 shows the corresponding HuggingFace repositories for the models used in the experiment.

F.3 WikiData Taxonomy

We extract the WikiTax taxonomy using the edges *subclass of*, *instances of* and *subproperty of* (Wikidata identifiers P279, P31 and P1647)[8].

[8] https://github.com/nichtich/wikidata-taxonomy

F.4 Correlation Coefficients

We use rank correlation since we do not expect a linear relationship between metric scores, as we expect their rankings to align rather than their absolute values.

G Extended Results

G.1 Extrinsic Evaluation

Fig. 14 shows scatter plots and a fitted exponential on the space of the collected metric scores of taxonomies mutated during the experiments. CSC and NLIV can, to an extent, predict downstream F1, as indicated by the R^2. While RaTE shows a positive tendency, there is higher variance and it does not work for DBP without tuning as shown in Fig. 5. The scatter plot for SP clearly shows its negative association with F1.

G.2 Stability of Metrics

Fig. 11 shows the violin plots for perturbed vs. mutated taxonomies now with RaTE and SP included.

G.3 Metrics

For intuition, Fig. 12 and Fig. 13 show scatterplots of the metrics for the collected samples of mutated taxonomies. Fig. 14 shows the downstream scores vs. downstream F1, including the fitted exponential curves with their parameterizations.

G.4 Approximating Parent-Child Likelihoods

Tab. 7 shows the performance on the binary classification of parent-child edges for all datasets. Fig. 15 shows the performance given the taxonomic distance between the potential parent-child edges in the ground truth taxonomy.

LLMs for NLI In preliminary experiments using a setting similar to Fig. 15, prompting `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) for NLI yielded comparable performance to NLI-trained models, and GPT-4o (OpenAI et al., 2024) was only slightly better. We did not consider these improvements worth the extra compute. Obtaining continuous probabilities from LLMs would also require additional calibration, which we avoided to keep the method simple and general.

Dataset	All		Non-Leaf	
	Number samples	Number degradations	Number samples	Number degradations
MeSH	250	50	250	50
CookBook	500	100	500	100
WikiTax	500	100	500	100
SemEval-Food	500	100	500	100
SemEval-Verb	250	50	250	50

Table 5: The number of samples and degradations per taxonomy for the empirical validation.

Model Name	Huggingface Link	Publication
BART-L	facebook/bart-large-mnli	Lewis et al. (2020)
BGE-M3	BAAI/bge-m3	Chen et al. (2024)
DeBERTa-L	microsoft/deberta-large-mnli	He et al. (2021)
DeBERTa-ML	MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	He et al. (2021); Laurer et al. (2024)
MiniLM-L6	sentence-transformers/all-MiniLM-L6-v2	Reimers and Gurevych (2019)
ML-E5-L	intfloat/multilingual-e5-large-instruct	Wang et al. (2024)
Para-MPNet	sentence-transformers/paraphrase-multilingual-mpnet-base-v2	Reimers and Gurevych (2019)
RoBERTa-L	FacebookAI/roberta-large-mnli	Liu et al. (2019)

Table 6: Huggingface identifiers and links for model abbreviations along with their corresponding publications.

Dataset	NLIV-Strong				NLIV-Weak			
	ROC-AUC	Precision	Recall	F1	ROC-AUC	Precision	Recall	F1
With concept descriptions								
MeSH	0.84	0.85	0.50	0.63	0.82	0.48	0.98	0.65
CookBook	0.90	0.92	0.68	0.78	0.87	0.58	0.95	0.72
WikiTax	0.76	0.78	0.59	0.67	0.78	0.57	0.92	0.70
SemEval-Food	0.84	0.79	0.65	0.72	0.81	0.53	0.99	0.69
SemEval-Verb	0.70	0.56	0.67	0.61	0.66	0.43	0.99	0.60
With concept names (no descriptions)								
SemEval-Food	0.84	0.79	0.65	0.72	0.81	0.53	0.99	0.69
MeSH	0.84	0.84	0.52	0.64	0.84	0.47	1.00	0.64

Table 7: Performance on the binary classification of true or false parent-child edges. The lower table shows experiments in simply using concept names instead of descriptions. Performance is preserved in cases where concept descriptions are unavailable.

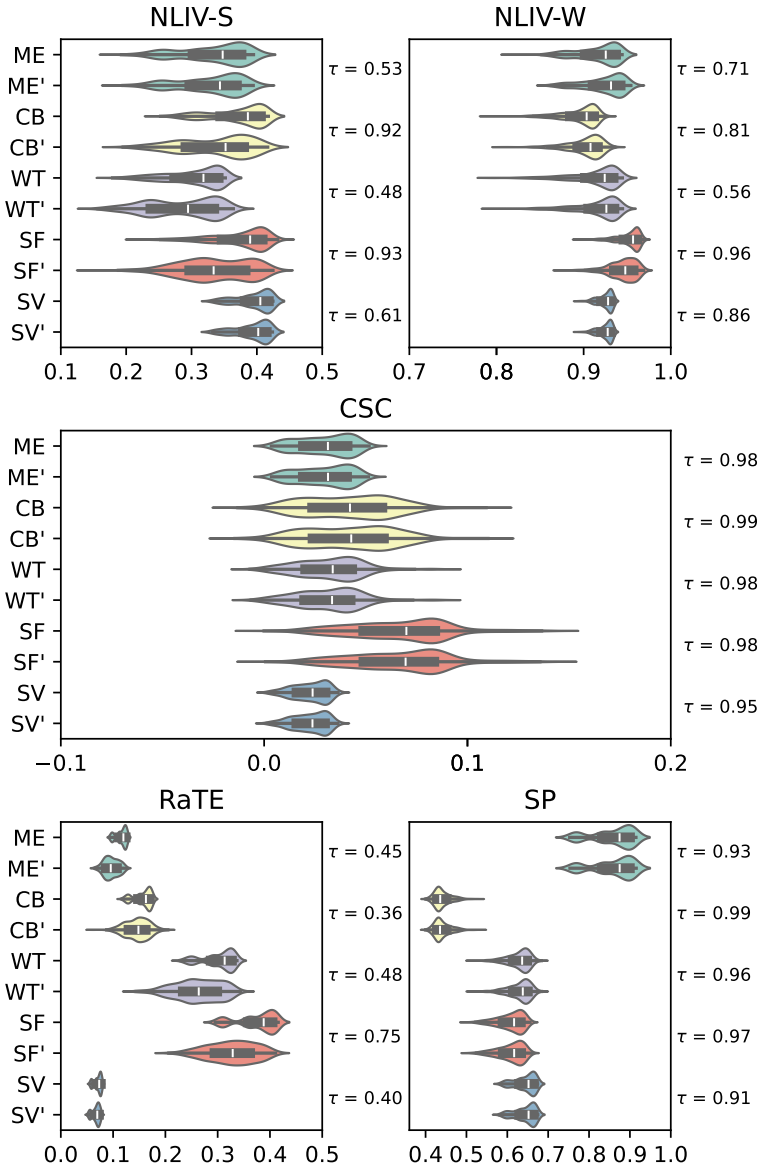
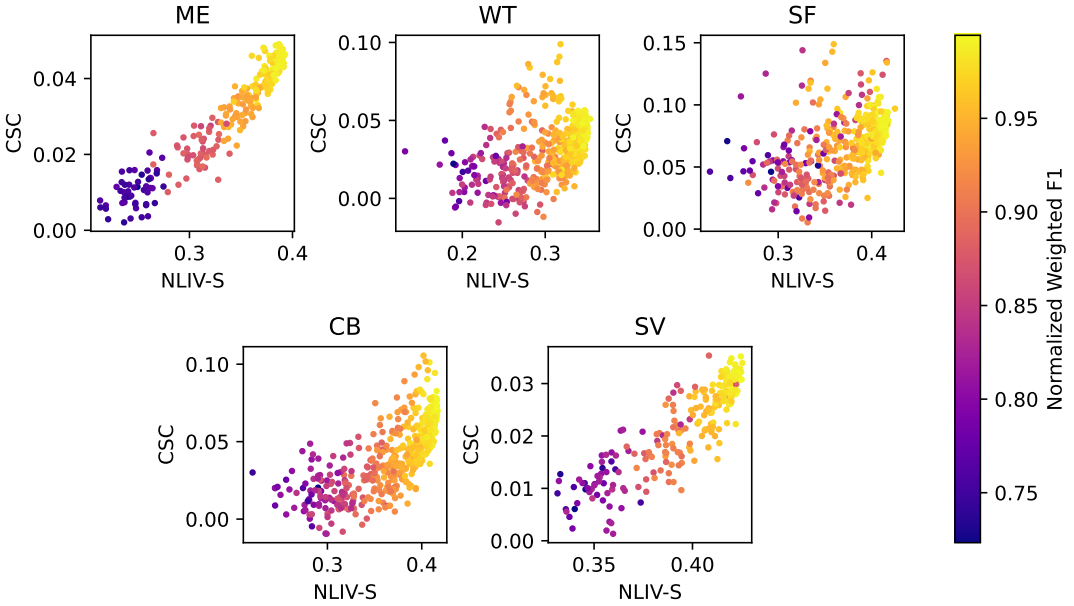
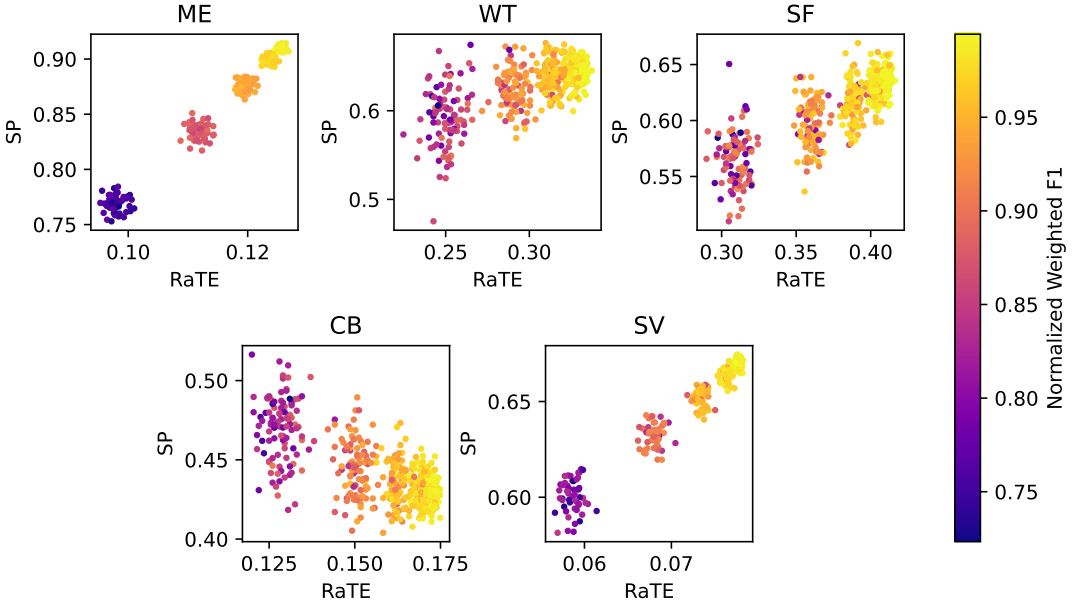


Figure 11: Violin plots showing the distribution of our scores when replacing concept names with synonyms using WordNet. This is the full plot including RaTE and SP metrics. The right hand side shows the Kendall's τ between the two distributions as a measure of their monotonic association. The distributions of perturbed taxonomies are indicated with primes (e.g., CB$'$).

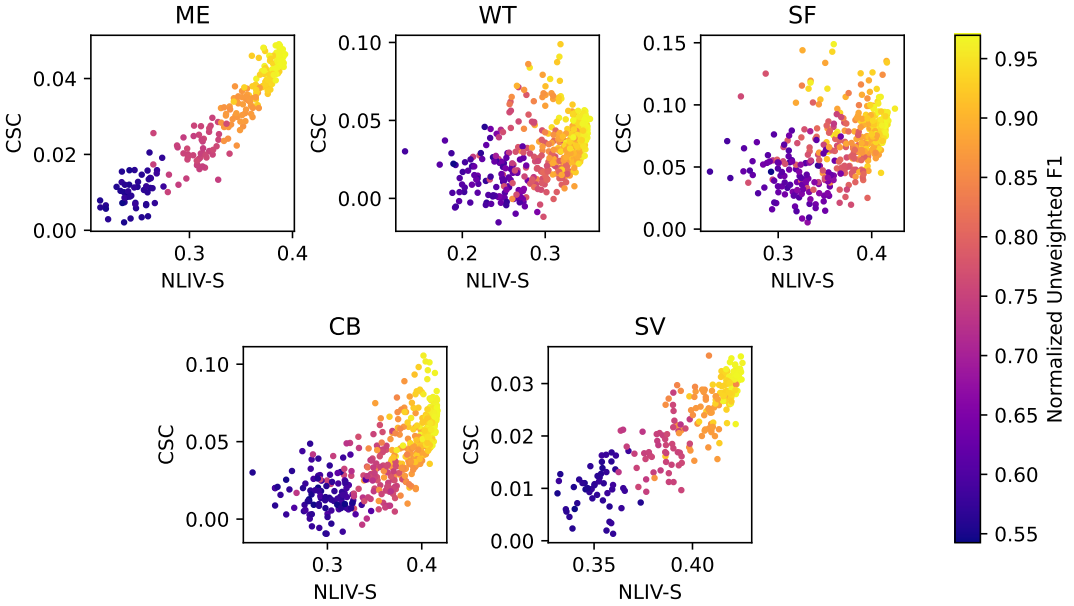


(a) Our metrics

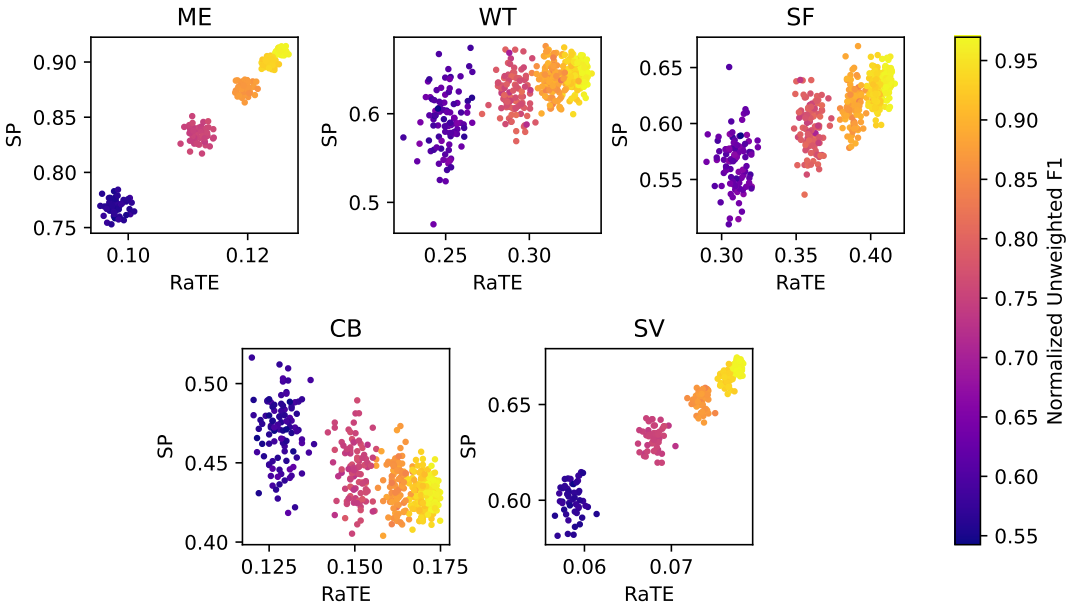


(b) Existing metrics

Figure 12: Scatter plots for the intrinsic evaluation samples with the color indicating the weighted F1 score against the ground truth taxonomy.



(a) Our metrics



(b) Existing metrics

Figure 13: Scatter plots for the intrinsic evaluation samples with the color indicating the unweighted F1 score against the ground truth taxonomy.

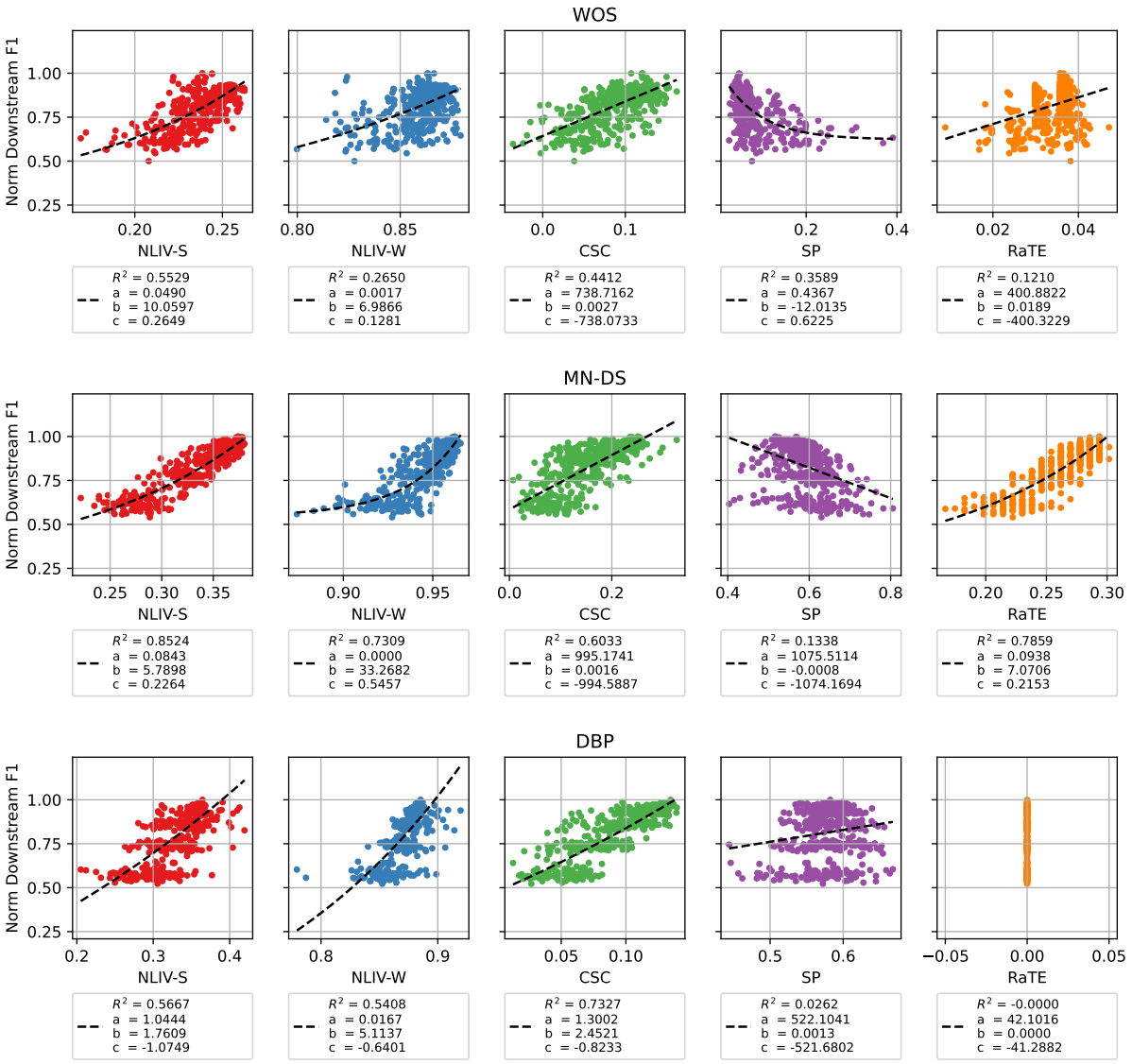


Figure 14: Scatter plots of metric scores vs. F1 on the downstream task including fitted exponentials of form $y = a \cdot e^{bx} + c$.

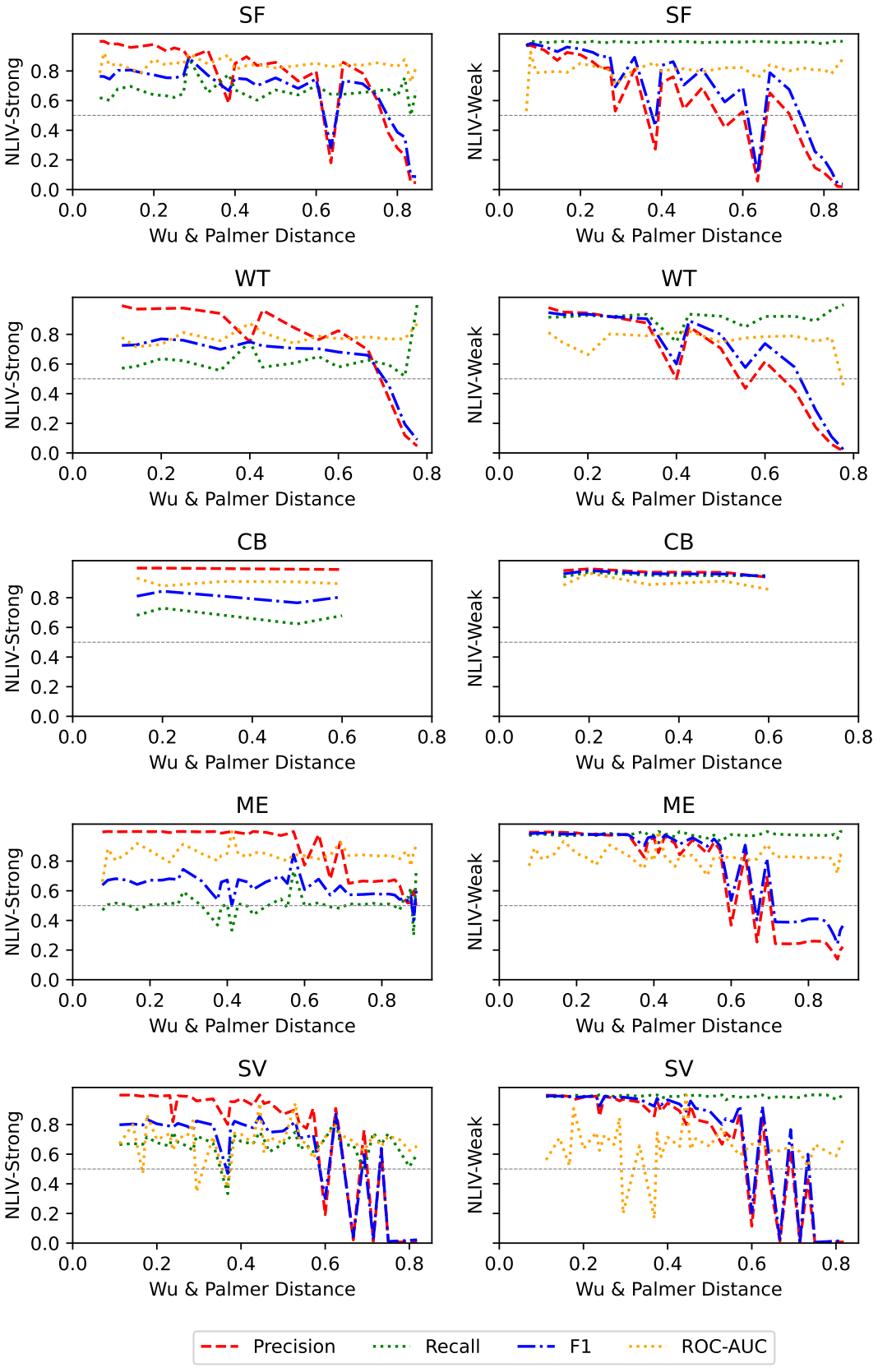


Figure 15: Performance of NLI for estimating edge probabilities measured in binary F1, precision and recall and ROC-AUC by moving over windows along the WPS of the concepts pairs of the edge. We randomly sample $|\mathcal{E}|$ negative pairs by randomly combining concepts without any common ancestor (except the root).