Chem-FINESE: Validating Fine-Grained Few-shot Entity Extraction through Text Reconstruction

Qingyun Wang, Zixuan Zhang, Hongxiang Li, Xuan Liu,
Jiawei Han, Huimin Zhao, Heng Ji
University of Illinois at Urbana-Champaign
{qingyun4,zixuan11,hanj,zhao5,hengji}@illinois.edu

Abstract

Fine-grained few-shot entity extraction in the chemical domain faces two unique challenges. First, compared with entity extraction tasks in the general domain, sentences from chemical papers usually contain more entities. Moreover, entity extraction models usually have difficulty extracting entities of long-tailed types. In this paper, we propose Chem-FINESE, a novel sequence-to-sequence (seq2seq) based few-shot entity extraction approach, to address these two challenges. Our Chem-FINESE has two components: a seq2seq entity extractor to extract named entities from the input sentence and a seq2seq self-validation module to reconstruct the original input sentence from extracted entities. Inspired by the fact that a good entity extraction system needs to extract entities faithfully, our new self-validation module leverages entity extraction results to reconstruct the original input sentence. Besides, we design a new contrastive loss to reduce excessive copying during the extraction process. Finally, we release ChemNER+, a new fine-grained chemical entity extraction dataset that is annotated by domain experts with the ChemNER schema. Experiments in few-shot settings with both ChemNER+ and CHEMET datasets show that our newly proposed framework has contributed up to 8.26% and 6.84% absolute F1-score gains respectively[1].

1 Introduction

Millions of scientific papers are published annually[2], resulting in an information overload (Van Noorden, 2014; Landhuis, 2016). Due to such an explosion of research directions, it is impossible for scientists to fully explore the landscape due to the limited reading ability of humans. Therefore, information extraction, especially entity extraction of fine-grained scientific entity types, becomes a crucial step to automatically catch up with the newest research findings in the chemical domain.

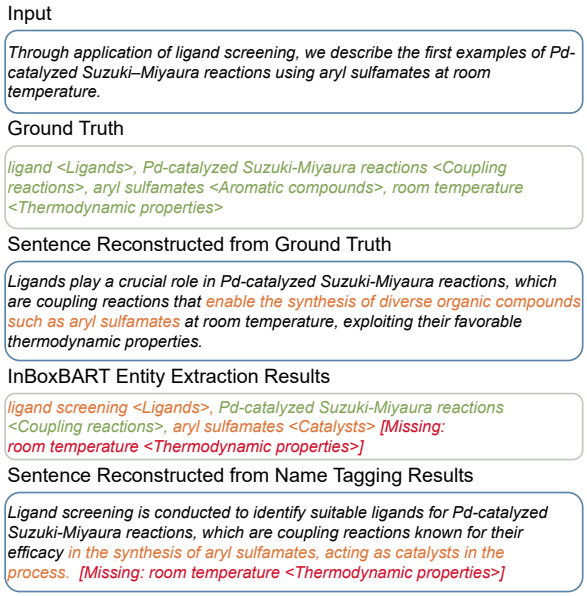

Figure 1: Comparison of sentence reconstruction results from ground truth and InBoXBART (Parmar et al., 2022). We highlight Complete Correct, *Missed Entity*, and *Partially Correct Prediction* with different color.

Despite such a pressing need, fine-grained entity extraction in the chemical domain presents three distinctive and non-trivial challenges. First, there are very few publicly available benchmarks with high-quality annotations on fine-grained chemical entity types. For example, ChemNER (Wang et al., 2021a) developed the first fine-grained chemistry entity extraction dataset. However, their dataset is not released publicly. To address this issue, we collaborate with domain experts to annotate ChemNER+, a new chemical entity extraction dataset based on the ChemNER ontology. Besides, we construct another new fine-grained entity extraction

[1]The programs, data, and resources are publicly available for research purposes at: https://github.com/EagleW/Chem-FINESE.

[2]https://esperr.github.io/pubmed-by-year/about.html

dataset based on an existing entity typing dataset CHEMET (Sun et al., 2021).

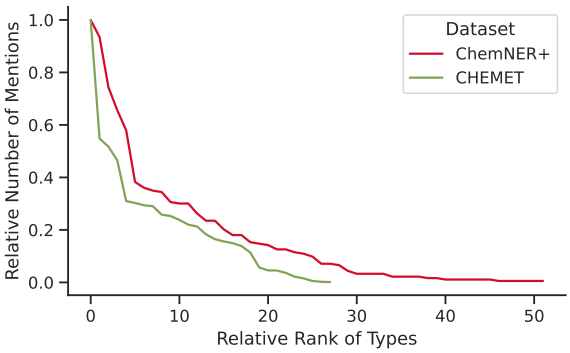

Figure 2: Type distributions for the training sets of ChemNER+ and CHEMET datasets. The Y-axis represents the number of mentions normalized by the mentions of the most frequent type. The X-axis represents the rank of types.

In addition, current entity extraction systems in few-shot settings face two main problems: *missing mentions* and *incorrect long-tail predictions*. One primary reason for missing mentions is that the sentences in scientific papers typically cover more entities than sentences in the general domain. For example, there are 3.1 entities per sentence in our ChemNER+ dataset, which is much higher than the 1.5 entities in the general domain dataset CONLL2003 (Tjong Kim Sang and De Meulder, 2003). As a result, it is more difficult for entity extraction models to cover all mentions in the input sentences. As shown in Figure 1, since the input has already included four chemical entities, InBoXBART model (Parmar et al., 2022) completely misses the entity *"room temperature"*.

Furthermore, entity distributions in the chemical domain are highly imbalanced. As shown in Figure 2, we observe that the entity type distributions of ChemNER+ and CHEMET exhibit similar long-tail patterns. In few-shot settings, entities with long-tail types are extremely difficult to extract due to insufficient training examples. For example, as shown in Figure 1, InBoXBART mistakenly predicts the entity *"aryl sulfamates"* as *catalyst*, because its type has a frequency forty times lower than the predicted type (i.e., 4 vs 136). Moreover, the diverse representation nature of chemical entities—such as trade names, trivial names, and semi-systematic names (e.g., THF, iPrMgCl, 8-phenyl ring)—makes it even harder for models to generalize on these long-tail entities.

To address these challenges, we propose a novel **Chem**ical **FIN**e-grained **E**ntity extraction with **SE**lf-validation (Chem-FINESE). Specifically, our Chem-FINESE has two parts: a seq2seq entity extractor to extract named entities from the input sentence and a seq2seq self-validation module to reconstruct the original input sentence based on the extracted entities. First, we employ a seq2seq model to extract entities from the input sentence, since it does not require any task-specific component and explicit negative training examples (Giorgi et al., 2022). We generate the entity extraction results as a concatenation of pairs, each consisting of an entity mention and its corresponding type, as shown in Figure 1.

One critical issue for seq2seq entity extraction is that the language model tends to miss important entities or excessively copy original input. For example, the seq2seq entity extraction results missed the type *thermodynamic properties* and generated *"ligand screening"* in Figure 1. However, the goal of information extraction is to provide factual information and knowledge comprehensively. In other words, *if the model extracts knowledge precisely, readers should be able to faithfully reconstruct the original sentence using the extraction results*. Inspired by such a goal, to evaluate whether the seq2seq entity extractor has faithfully extracted important information, we propose a novel seq2seq self-validation module to reconstruct the original sentences based on entity extraction results. As shown in Figure 1, the sentence reconstructed from the ground truth is closer to the original input than the sentence reconstructed from entity extraction results, which misses the reaction condition and introduces additional information that treated the *"aryl sulfamates"* as *catalysts*. Additionally, we introduce a new entity decoder contrastive loss to control the mention spans. We treat text spans containing entity mentions as hard negatives. For instance, given the ground truth entity *"aryl sulfamates"*, we will treat *"aryl sulfamates at room temperature"* as a hard negative.

Our extensive experiments demonstrate that our proposed framework significantly outperforms our baseline model by up to 8.26% and 6.84% absolute F1-score gains on ChemNER+ and CHEMET datasets respectively. Our analysis also shows that Chem-FINESE can effectively learn to select correct mentions and improve long-tail entity type performance. To evaluate the generalization ability of our proposed method, we also evaluate our framework on CrossNER (Liu et al., 2021), which

is based on Wikipedia. Our Chem-FINESE still outperforms other baselines in all five domains.

Our contributions are threefold:

1. We propose two few-shot chemical fine-grained entity extraction datasets, based on human-annotated ChemNER+ and CHEMET.

2. We propose a new framework to address the mention coverage and long-tailed entity type problems in chemical fine-grained entity extraction tasks through a novel self-validation module and a new entity extractor decoder contrastive objective. Our model does not require any external knowledge or domain adaptive pretraining.

3. Our extensive experiments on both chemical few-shot fine-grained datasets and the CrossNER dataset justify the superiority of our Chem-FINESE model.

2 Task Formulation

Following Giorgi et al. (2022), we formulate entity extraction as a sequence-to-sequence (seq2seq) generation task by taking a source document $\mathcal{S}$ as input. The model generates output $\mathcal{Y}$, a text consisting of a concatenation of n fine-grained chemical entities $E_1, E_2, ..., E_n$. Each mention E_i includes the mention μ_i in the source document $\mathcal{S}$ and its entity type $\rho_i \in \mathcal{P}$, where $\mathcal{P}$ is a set containing all entity types. Specifically, we propose the following output linearization schema: given the input $\mathcal{S}$, the output is $\mathcal{Y} = \mu_1 < \rho_1 >, \mu_2 < \rho_2 >, ..., \mu_n < \rho_n >$. We further illustrated this with an example:

$\mathcal{S}$: Through application of ligand screening, we describe the first examples of Pd-catalyzed Suzuki–Miyaura reactions using aryl sulfamates at room temperature.

$\mathcal{Y}$: ligand <Ligands>, Pd-catalyzed Suzuki–Miyaura reactions <Coupling reactions>, aryl sulfamates <Aromatic compounds>, room temperature <Thermodynamic properties>

3 Method

3.1 Model Architecture

The overall framework is illustrated in Figure 3. Given the source document S, we first use a seq2seq model to extract fine-grained chemical entities. Then, we propose a new *self-validation module* to reconstruct the original input based on entity extraction results. Finally, we introduce a new *entity decoder contrastive loss* to reduce excessive copying. The entire model is trained with a combination of the supervised loss, the reconstruction loss, and the entity decoder contrastive loss.

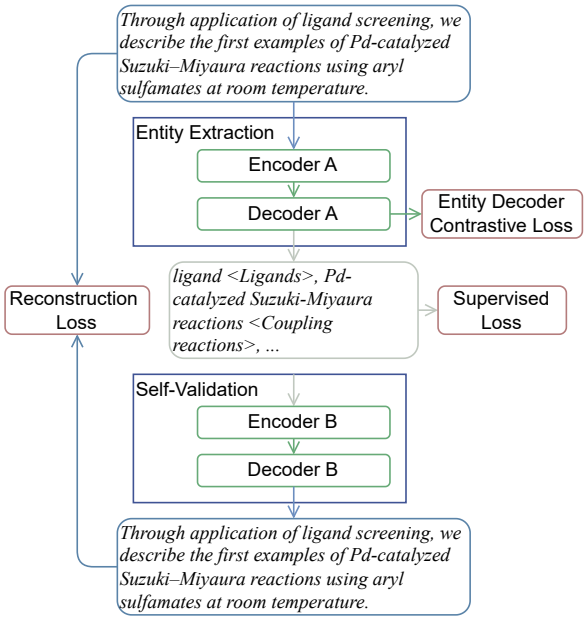

Figure 3: Architecture overview. We use the example in Figure 1 as a walking-through example.

3.2 Entity Extraction Module

Our entity extraction module follows a seq2seq setup (Yan et al., 2021; Giorgi et al., 2022). Formally, we use the state-of-the-art coarse-grained chemical entity extractor InBoXBART (Parmar et al., 2022) as the backbone. We model the conditional probability of extracting entities from source sequence $\mathcal{S}$ as

$$p(\mathcal{Y}|\mathcal{S}) = \prod_{t=1}^{T} p(y_t|\mathcal{S}, y_{<t}), \quad (1)$$

where the output $\mathcal{Y}$ has a length of T, and y_t is the predicted token at time t in the output $\mathcal{Y}$.

We supervise the entity extraction using the standard cross-entropy loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log p(y_t|\mathcal{S}, y_{<t}). \quad (2)$$

3.3 Self-validation Module

Since a good information extraction system needs to extract entities faithfully, we propose a self-validation module to reconstruct the original sentence from the extracted entities to check whether the model overlooks any entities. Different from previous dual learning architectures (Iovine et al.,

2022), which use dual cycles or reinforcement learning to provide feedback, we use Gumbel-softmax (GS) estimator (Jang et al., 2017) to avoid the non-differentiable issue in explicit decoding. Specifically, based on InBoXBART (Parmar et al., 2022), we first pretrain a seq2seq self-validation module that takes in the entity extraction results $\mathcal{Y}$ and generates a reconstructed sentence $\hat{\mathcal{S}}$. We use our training set to pretrain the self-validation module. We fix the weight of the self-validation module after pretraining. In the training stage, the input embedding $\mathbf{H}_t$ of the self-validation module is given by:

$$\mathbf{H}_t = \mathsf{GS}\left(p\left(y_t|\mathcal{S}, y_{<t}\right)\right) \cdot \mathbf{E}_v, \tag{3}$$

where $\mathbf{E}_v$ is the vocabulary embedding matrix and GS is the Gumbel-softmax estimator. The total input embeddings for the self-reconstruction model is $\mathbf{H} = [\mathbf{H}_1; \mathbf{H}_2; ...; \mathbf{H}_T]$.

The reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = -\sum_{\hat{t}=1}^{\hat{T}} \log p(\hat{s}_{\hat{t}}|\mathbf{H}, \hat{s}_{<\hat{t}}), \tag{4}$$

where the reconstructed sentence $\hat{\mathcal{S}}$ has a length of $\hat{T}$, and $\hat{s}_{\hat{t}}$ is the predicted token at time $\hat{t}$ in $\hat{\mathcal{S}}$.

3.4 Contrastive Entity Decoding Module

Entity extraction datasets in the scientific domain usually contain more entities for each sentence. From the initial experiments, we found that the entity extraction module tends to generate incorrect mentions by associating it with unrelated contexts to help the reconstruction of the self-validation module. For example, given the example in Figure 1, the baseline model generates *"ligand screening"* instead of *"ligand"*. Therefore, we introduce a new decoding contrastive loss inspired by Wang et al. (2023a) to suppress excessive copying. We construct negative samples by combining mentions with surrounding unrelated contexts. For example, we will consider *"ligand screening, we describe the first examples"* as a negative of entity *"ligand"*. We treat the original mention type pairs as the ground truth and maximize their probability with InfoNCE loss (Oord et al., 2018):

$$\begin{aligned}
\mathcal{L}_{\text{cl}} &= \frac{\exp\left(x^+/\tau\right)}{\sum_i \exp\left(x_i^-/\tau\right) + \exp\left(x^+/\tau\right)}, \\
x^+ &= \sigma(\text{Avg}(\mathbf{W}_x\bar{\mathbf{H}}^+ + \mathbf{b}_x)), \\
x_i^- &= \sigma(\text{Avg}(\mathbf{W}_x\bar{\mathbf{H}}_i^- + \mathbf{b}_x)),
\end{aligned} \tag{5}$$

where $\bar{\mathbf{H}}^+$ and $\bar{\mathbf{H}}_i^-$ are decoder hidden states from the positive and i-th negative samples, $\mathbf{W}_x$ is a learnable parameter, τ is the temperature, and $\text{Avg}(*)$ denotes the average pooling function.

3.5 Training Objective

We jointly optimize the cross-entropy loss, reconstruction loss, and entity decoder contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \alpha\mathcal{L}_{\text{recon}} + \beta\mathcal{L}_{\text{cl}}, \tag{6}$$

where α, β are hyperparameters that control the weights of the reconstruction loss and contrastive loss respectively.

Dataset	Split	#Pair	$\overline{\#\textbf{Token}}$	$\overline{\#\textbf{Entity}}$
ChemNER+	Train	542	32.9	3.10
	Valid	100	39.9	4.57
	Test	100	39.4	4.61
CHEMET	Train	6,561	37.8	1.57
	Valid	520	31.6	2.15
	Test	663	36.6	1.95

Table 1: Statistics of our dataset. $\overline{\#\textbf{Token}}$ denotes average number of words per sentence. $\overline{\#\textbf{Entity}}$ denotes average number of entities per sentence.

4 Benchmark Dataset

4.1 Dataset Creation

ChemNER+ Dataset. Since the annotation of ChemNER dataset is not fully available online, we decide to create our own dataset, ChemNER+, based on available sentences from ChemNER (Wang et al., 2021a) dataset. Following the schema of ChemNER, we ask two Chemistry Ph.D. students to annotate a new dataset, covering 59 fine-grained chemistry types with 742 sentences[3].

CHEMET Dataset. We construct a new fine-grained entity extraction dataset based on CHEMET (Sun et al., 2021). For any entity in the training set that overlaps with the validation and testing sets, we replace its multi-labels with the most frequent types that appear in the validation and testing sets. For other entities, we replace the remaining types with their most frequent types that appeared in the training set. We merge the entity types with the same subcategory name in CHEMET (Sun et al., 2021). The final dataset consists of 30 fine-grained organic chemical types.

Table 1 shows the detailed data statistics.

[3]Human annotation details are in Appendix E.

k-shot	6	9	12	15	18
RoBERTa	8.09	7.98	8.00	16.22	7.94
PubMedBERT	5.48	5.12	5.77	5.46	5.88
ScholarBERT	23.96	29.82	27.65	31.48	32.76
NNShot	0.99	1.43	2.39	1.61	2.45
StructShot	0.86	1.32	2.27	1.62	2.47
InBoXBART	26.23	27.89	28.83	33.64	30.39
+ Valid	32.40	31.13	33.64	35.31	36.44
+ Valid + CL	**33.11**	**32.75**	**34.75**	**37.89**	**38.65**

Table 2: micro-F1 (%) scores for ChemNER+ with few-shot settings. *Valid* is a model with a self-validation module. *CL* is a model with a decoder contrastive loss.

k-shot	6	9	12	15	18
RoBERTa	4.91	4.16	4.79	4.83	4.81
PubMedBERT	4.07	4.67	3.87	4.47	3.96
ScholarBERT	17.00	33.63	29.65	29.72	32.52
NNShot	4.23	4.03	4.14	5.27	4.76
StructShot	4.15	4.00	4.19	5.21	4.79
InBoXBART	29.93	29.57	31.76	36.16	37.52
+ Valid	32.74	34.09	33.30	**40.81**	38.37
+ Valid + CL	**33.81**	**36.41**	**36.11**	40.52	**39.94**

Table 3: micro-F1 (%) scores for CHEMET with few-shot settings.

4.2 Few-shot Setup

For each dataset, we randomly sample a subset based on the frequency of each type class. Specifically, given a dataset, we first set the number of maximum entity mentions k for the most frequent entity type in the dataset. We then randomly sample other types and ensure that the distribution of each type remains the same as in the original dataset. We choose the values $6, 9, 12, 15, 18$ as the potential maximum entity mentions for k. The ChemNER+ and CHEMET few-shot datasets contain 52 and 28 types respectively.

5 Experiments

5.1 Baselines

We compare our model with **(1) state-of-the-art pretrained encoder-based models** including RoBERTa (Liu et al., 2019) and models with domain adaptive training, such as PubMedBERT (Gu et al., 2021) and ScholarBERT (Hong et al., 2023). We then compare our model with the **(2) few-shot baselines**, including NNShot and StructShot (Yang and Katiyar, 2020) based on RoBERTa-base. Since we use **InBoXBART** (Parmar et al., 2022) as our backbone, we also include **(3) baselines for ablation**. The hyperparameters, training and evaluation details are presented in Appendix A.

5.2 Overall Performance

Tables 2, 3 show that our models outperform baselines for few-shot settings by a large margin. Compared to the best pretrained encoder-based ScholarBERT, pretrained on 221B tokens of scientific documents, seq2seq models generally achieve higher performance in low-resource settings with fewer parameters, as shown in Table 11. We also observe that both NNshot and StructShot perform worse than their original baseline. At a closer look, we find that both methods miss many entities and mislabel unrelated phrases as entities. The primary reasons for this are twofold: first, the chemical domain's entity mentions are more diverse and may only appear in the testing set; second, there are significantly more potential entity types than in traditional entity extraction tasks. Therefore, the two baselines cannot effectively utilize the nearest neighbor information and perform worse than our proposed methods. These results demonstrate that seq2seq models have a better generalization ability in few-shot settings.

k-shot	6	9	12	15	18
InBoXBART	36.96	38.22	38.34	47.91	42.84
+ Valid	45.07	**45.28**	41.56	48.15	46.15
+ Valid + CL	**45.58**	44.03	**45.25**	**51.68**	**47.88**

Table 4: Mention micro-F1 (%) scores for ChemNER+ with few-shot settings.

k-shot	6	9	12	15	18
InBoXBART	46.74	42.07	44.32	47.58	52.90
+ Valid	47.87	46.01	44.18	50.55	50.50
+ Valid + CL	**48.96**	**49.83**	**47.03**	**50.61**	**54.10**

Table 5: Mention micro-F1 (%) scores for CHEMET with few-shot settings.

Additionally, the self-validation variants significantly outperform the baseline InBoXBART, showing the benefit of the self-validation module in capturing mentions. Moreover, our self-validation module can effectively enhance the performance of the entity extraction module in extremely low-resource settings. In 6-shot scenarios for both ChemNER+ and CHEMET datasets, our model achieves impressive performance compared to ScholarBERT, which further verifies the effectiveness of the self-validation module. Finally, adding decoder contrastive loss helps the model perform significantly better in Table 2, suggesting

that contrastive learning further helps the mention extraction quality by reducing excessive copying. Interestingly, we observe that decoder contrastive learning improves less in Table 3 than in Table 2, because the CHEMET contains fewer entities per sentence compared to the ChemNER+.

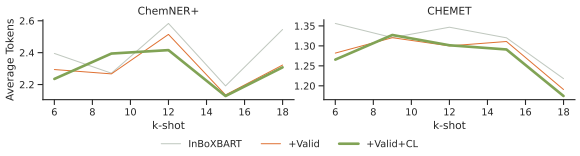

Figure 4: Average tokens in each mention for ChemNER+ and CHEMET datasets with few-shot settings.

Performance of Mention Extraction. We calculate the mention F1 scores in Tables 4 and 5. In addition, we also test a fully unsupervised mention extraction based on AMR-Parser (Fernandez Astudillo et al., 2020)[4]. The F1-scores are 38.22 and 45.33 for ChemNER+ and CHEMET, respectively. These results imply that the self-validation model generally improves the mention extraction accuracy. Moreover, adding decoder contrastive loss generally further bolsters the mention F1 score by reducing the number of tokens that appear in each mention, as shown in Figure 4.

k-shot	6	9	12	15	18
RoBERTa	2.04	2.05	2.05	0.00	2.05
PubMedBERT	2.05	0.00	0.00	2.13	0.00
ScholarBERT	0.00	9.28	4.71	0.00	6.90
InBoXBART	8.33	11.36	15.22	17.14	7.69
+ Valid	10.81	12.24	10.26	9.76	23.81
+ Valid + CL	**26.19**	**23.91**	**23.26**	**19.05**	**25.00**

Table 6: micro-F1 (%) scores for long-tail entity types ChemNER+ with few-shot settings.

Performance of Long-tail Entity. To evaluate the performance of long-tail entities, we first rank entity types by their frequency. We then select the entity types that appear in the lower 50% and calculate the F1 scores of those types[5]. The results are in Tables 6 and 7. Notably, our proposed methods greatly outperform the encoder-based baselines. Both the self-verification module and the decoder contrastive loss aid the entity extraction module in focusing on long-tail entities by creating a more balanced distribution of entity types. The major reason for the relatively low performance in Table 7 is that

[4]Implementation details are in Appendix A.

[5]Entity frequency and selected types are in Appendix B.

k-shot	6	9	12	15	18
RoBERTa	0.00	0.00	0.00	0.00	0.00
PubMedBERT	0.00	0.00	0.00	0.00	0.00
ScholarBERT	0.00	0.00	0.00	0.00	0.00
InBoXBART	4.90	7.55	4.55	5.05	12.26
+ Valid	**8.72**	**13.10**	4.55	**16.96**	20.83
+ Valid + CL	7.07	11.32	**8.33**	5.15	**23.01**

Table 7: micro-F1 (%) scores for long-tail entity types CHEMET with few-shot settings. The encoder-based models fail to extract long-tail entity types for all few-shot settings. Compared to encoder-based models, seq2seq models can utilize label semantics in the generation procedure. Therefore, encoder-based models require more training data under few-shot settings.

the differences between the types in CHEMET are not significant. The relatively stable performance of our model in Table 6 across increasing few-shot examples indicates that our model achieves satisfactory performance for long-tail entities, even with a limited training sample.

6 Analysis

6.1 Qualitative Analysis

Table 8 shows two typical examples from the 18-shot ChemNER+ dataset that illustrate how incorporating a self-validation module and decoder contrastive loss can improve the mention coverage and long-tail entity performance.

In the first example, the InBoXBART baseline fails to identify both *"cyclophanes"* and *"polycycles"*, probably because the input sentence contains too many entities. With the help of the self-validation module, the InBoXBART+Valid model successfully captures the first entity *"cyclophanes"*. However, it still cannot recognize *"polycycles"*. Additionally, both the baseline and the InBoXBART+Valid model mistakenly treat the entity *"Suzuki cross-coupling and metathesis"* and the entity *"metathesis"*, because those models excessively copy from the original sentence. In contrast, by adding the decoder contrastive loss, which uses the mentions with surrounding unrelated contexts as negatives, the model successfully separates the entity *"Suzuki cross-coupling and metathesis"* from the entity *"metathesis"*.

In the second example, both the baseline and the InBoXBART+Valid model predict a very long text span that treats three entities as a single entity. They also fail to capture *"asymmetric catalysis"* and *"highly enantioselective process"* as entities because their types have low frequency in the train-

ing set. With the help of decoder contrastive loss, the model reduces the excessive copying of the entity extraction module while trying to capture important entities as accurately as possible. Therefore, the model successfully classifies *"asymmetric catalysis"* as *Catalysis* correctly and also predicts *"enantioselective process"* as an entity.

6.2 Compatible with Other Few-shot Datasets?

CrossNER Dataset. In the above experiments, we focus on the few-shot settings for chemical papers and prove the effectiveness of our proposed framework. To evaluate the generalization ability of our proposed framework on other domains, we conduct experiments on the CrossNER dataset (Liu et al., 2021). The detailed statistics are in Table 9. We remove sentences without any entity. Because the CrossNER dataset is based on Wikipedia articles, we choose RoBERTa and ScholarBERT as encoder-based baselines. Additionally, we select BART-base (Lewis et al., 2020) as the backbone for our ablation variations.

Results. As shown in Table 10, our model consistently produces the best F1 scores across all five domains of CrossNER without any external knowledge or domain adaptive pretraining. We observe that the model achieves the largest gain for the AI domain and the smallest gain for the politics domain. The major reason behind this is that AI domain contains the most informative entity types, which cover the key points of the sentence, including *algorithm*, *task*, etc. On the contrary, the politics domain contains many names of *politicians* and *locations*, which require background knowledge for the self-verification module to identify.

6.3 Remaining Challenges

Misleading Subwords. We observe that the mention text can sometimes mislead the type predictions, especially if the type contains a subword from the mention. As a result, the model fails to identify the type correctly. For example, given the mention *"unnatural amino acid derivatives"*, our model focuses on the word "acid" and predicts the entity to be *Organic acids* instead of *Organonitrogen compounds*. The potential reason behind this is that the BART model incorrectly associates the *"acid"* in the mention with *Organic acids*. Such type errors might be incorporated into the decoder contrastive learning as additional hard negatives.

Fine-grained Type Classification. The model tends to predict generic entity types instead of more fine-grained entity types. For instance, the model predicts the mention *"Cs2CO3"* as *Inorganic compounds* instead of *Inorganic carbon compounds*. This issue might come from annotation ambiguity in the training set. Additionally, the model predicts types that are not in the predefined ontology. For instance, the model labels *"GK"* as *Genecyclic compounds* instead of *Enzymes*. This error can possibly be solved by constraint decoding.

7 Related Work

Scientific Entity Extraction. Entity extraction for scientific papers has been widely exploited in the biomedical domain (Nguyen et al., 2022; Labrak et al., 2023; Cao et al., 2023; Li et al., 2023b; Hiebel et al., 2023) and the computer science domain (Luan et al., 2018; Jain et al., 2020; Viswanathan et al., 2021; Shen et al., 2021; Ye et al., 2022; Jeong and Kim, 2022; Hong et al., 2023). Despite this, fine-grained scientific entity extraction (Wang et al., 2021a) in the chemical domain receives less attention due to the scarcity of benchmark resources. Most benchmarks in the chemical (Krallinger et al., 2015; Kim et al., 2015) only provide coarse-grained entity types. In this paper, we address this problem by releasing two new datasets for chemical fine-grained entity extraction based on the ChemNER schema (Wang et al., 2021a) and CHEMET dataset (Sun et al., 2021).

Few-shot Entity Extraction. Few-shot learning attracts growing interest, especially for low-resource domains. Previous improvements for few-shot learning can be divided into several categories: domain-adaptive training by training the model in the same or similar domains (Liu et al., 2021; Oh et al., 2022), prototype learning by learning entity type prototypes (Ji et al., 2022; Oh et al., 2022; Ma et al., 2023), prompt-based methods (Lee et al., 2022; Xu et al., 2023; Nookala et al., 2023; Yang et al., 2023; Chen et al., 2023b), data-augmentation (Cai et al., 2023; Ghosh et al., 2023), code generation (Li et al., 2023a), meta-learning (de Lichy et al., 2021; Li et al., 2022; Ma et al., 2022), knowledge distillation (Wang et al., 2021c; Chen et al., 2023a), contrastive learning (Das et al., 2022), and external knowledge including label definitions (Wang et al., 2021b), AMR graph (Zhang et al., 2021), and background

InBoXBART	Several *cyclophanes*, *polycycles*, ... have been synthesized by employing a combination of *Suzuki cross-coupling and metathesis* Coupling reactions.
+Valid	Several *cyclophanes* Heterocyclic compounds, *polycycles*, ... have been synthesized by employing a combination of *Suzuki cross-coupling and metathesis* Organic reactions.
+Valid+CL	Several *cyclophanes* Heterocyclic compounds, *polycycles* Biomolecules, ... have been synthesized by employing a combination of Suzuki cross-coupling Coupling reactions and *metathesis* Chemical properties.
Ground Truth	Several cyclophanes Aromatic compounds, polycycles Organic polymers, ... have been synthesized by employing a combination of Suzuki cross-coupling Coupling reactions and metathesis Substitution reactions.
InBoXBART	... with the advantages of *asymmetric catalysis* (step and atom economy) in a rare example of an *enantioselective cross coupling of a racemic electrophile bearing an oxygen leaving group* Catalysis ... the identification of a *highly enantioselective process*.
+Valid	... with the advantages of *asymmetric catalysis* (step and atom economy) in a rare example of an *enantioselective cross coupling of a racemic electrophile bearing an oxygen leaving group* Organometallic compounds ... the identification of a *highly enantioselective process*
+Valid+CL	...with the advantages of asymmetric catalysis Catalysis (step and atom economy) in a rare example of an *enantioselective cross coupling* of a *racemic electrophile* bearing an oxygen leaving group Functional groups ... the identification of a highly enantioselective process Chemical properties.
Ground Truth	... with the advantages of asymmetric catalysis Catalysis (step and atom economy) in a rare example of an enantioselective cross coupling Coupling reactions of a racemic electrophile Organic compounds bearing an oxygen leaving group Functional groups ... the identification of a highly enantioselective process Catalysis.

Table 8: Examples showing how the self-validation module and entity decoder contrastive loss improves the model performance. We highlight Complete Correct, *Missed Entity*, and *Partially Correct Prediction* with different color. Compared to other baselines, our **+Valid+CL** successfully captures entities where other baselines miss.

Dom.	Train	Valid	Test	#Type	#Token	#Entity
AI	100	350	430	14	31.5	4.42
Lit.	99	400	416	12	37.6	5.39
Mus.	100	380	465	13	41.4	7.05
Pol.	200	541	651	9	43.5	6.46
Sci.	200	450	543	17	35.8	5.62

Table 9: Statistics of CrossNER. *Dom.* denotes the domain of the dataset.

Model	AI	Lit.	Mus.	Pol.	Sci.
RoBERTa	60.88	67.51	59.07	63.79	60.96
ScholarBERT	56.99	59.35	52.26	57.15	57.01
BART-base	59.20	66.90	62.78	67.99	62.18
+ Valid	61.84	67.97	60.94	67.22	62.40
+ Valid + CL	**62.48**	**68.22**	**63.39**	**68.03**	**62.87**

Table 10: F1 (%) scores for CrossNER.

knowledge (Lai et al., 2021). In contrast to these methods, our approach formulates the task in a text-to-text framework. In addition, we introduce a new simple but effective self-validation module, which achieves competitive performance without external knowledge or domain adaptive training.

Cycle Consistency. Cycle consistency, namely structural duality, leverages the symmetric structure of tasks to facilitate the learning process. It has emerged as an effective way to deal with low-resource tasks in natural language processing. First introduced in machine translation (He et al., 2016; Cheng et al., 2016; Lample et al., 2018; Mohiuddin and Joty, 2019; Xu et al., 2020) to deal with the scarcity of parallel data, cycle consistency has been expanded to other natural language processing tasks, including semantic parsing (Cao et al., 2019; Ye et al., 2019), natural language understanding (Su et al., 2019; Tseng et al., 2020; Su et al., 2020), and data-to-text generation (Dognin et al., 2020; Guo et al., 2020; Wang et al., 2023b). Recently, Iovine et al. (2022) successfully apply the cycle consistency to entity extraction by introducing an iterative two-stage cycle consistency training procedure. Despite these efforts, the non-differentiability of the intermediate text in the cycle remains unsolved, leading to the inability to propagate the loss through the cycle. To address this issue, Iovine et al. (2022) and Wang et al. (2023b) alternatively freeze one of the two models in two adjacent cycles. On the contrary, we introduce the gumbel-softmax estimator to avoid the non-differentiable issue. Additionally, we reduce the dual cycle training into end-to-end training to save time and computation resources.

8 Conclusion and Future Work

In this paper, we introduce a novel framework for chemical fine-grained entity extraction. Specifi-

cally, we target two unique challenges for few-shot fine-grained scientific entity extraction: mention coverage and long-tail entity extraction. We build a new self-validation module to automatically proofread the entity extraction results and a novel decoder contrastive loss to reduce excessive copying. Experimental results show that our proposed model achieves significant performance gains on two datasets: ChemNER+ and CHEMET. In the future, we plan to explore incorporating an external knowledge base to further improve the model's performance. Specifically, we plan to inject type definition into the representation to facilitate the entity extraction procedure. We will also continue exploring the use of constraint decoding to further improve entity extraction quality.

9 Limitations

9.1 Limitations of Data Collections

Both ChemNER+ and CHEMET are based on papers about Suzuki Coupling reactions from PubMed[6]. Our fine-grained entity extraction datasets are biased towards the topics and ontology provided by ChemNER+ and CHEMET. For example, CHEMET only focuses on the organic compounds. The number of available sentences is limited by the original dataset and our annotation efforts. We currently only focus on the English sentences. We only test our model on chemical papers (i.e., ChemNER+ and CHEMET) and Wikipedia (CrossNER). In the future, we aim to adapt our model for categories in other languages.

9.2 Limitations of System Performance

Our few-shot learning framework currently requires defining the entity ontology and few-shot examples before performing any training and testing. Therefore, due to patterns in the pretraining set, our model might produce mention types that don't align with our predefined ontology. For instance, it may generate *Cyclopentadienyl compounds* instead of the predefined type *Cyclopentadienyl complexes*. Furthermore, the pretrained model might emphasize language modeling over accurately identifying entire chemical phrases. For example, it might recognize *Pd* in the catalyst *Pd(OAC)2* simply as a *transition metal*.

[6] https://pubmed.ncbi.nlm.nih.gov/

Acknowledgement

This work is supported by the Molecule Maker Lab Institute: an AI research institute program supported by NSF under award No. 2019897 , and by DOE Center for Advanced Bioenergy and Bioproducts Innovation U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DESC0018420. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of, the National Science Foundation, the U.S. Department of Energy, and the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Jiong Cai, Shen Huang, Yong Jiang, Zeqi Tan, Pengjun Xie, and Kewei Tu. 2023. Graph propagation based data augmentation for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 110–118, Toronto, Canada. Association for Computational Linguistics.

Jiarun Cao, Niels Peek, Andrew Renehan, and Sophia Ananiadou. 2023. Gaussian distributed prototypical network for few-shot genomic variant detection. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 26–36, Toronto, Canada. Association for Computational Linguistics.

Ruisheng Cao, Su Zhu, Chen Liu, Jieyu Li, and Kai Yu. 2019. Semantic parsing with dual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 51–64, Florence, Italy. Association for Computational Linguistics.

Jiawei Chen, Yaojie Lu, Hongyu Lin, Jie Lou, Wei Jia, Dai Dai, Hua Wu, Boxi Cao, Xianpei Han, and Le Sun. 2023a. Learning in-context learning for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13661–13675, Toronto, Canada. Association for Computational Linguistics.

Yanru Chen, Yanan Zheng, and Zhilin Yang. 2023b. Prompt-based metric learning for few-shot NER. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7199–7212, Toronto, Canada. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised

learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany. Association for Computational Linguistics.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022. CONTaiNER: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Cyprien de Lichy, Hadrien Glaude, and William Campbell. 2021. Meta-learning for few-shot named entity recognition. In *Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing*, pages 44–58, Online. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Pierre Dognin, Igor Melnyk, Inkit Padhi, Cicero Nogueira dos Santos, and Payel Das. 2020. DualTKB: A Dual Learning Bridge between Text and Knowledge Base. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8605–8616, Online. Association for Computational Linguistics.

Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1001–1007, Online. Association for Computational Linguistics.

Alyson Gamble. 2017. Pubmed central (pmc). *The Charleston Advisor*, 19(2):48–54.

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, Ramaneswaran S, and Dinesh Manocha. 2023. ACLM: A selective-denoising based generative data augmentation approach for low-resource complex NER. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 104–125, Toronto, Canada. Association for Computational Linguistics.

John Giorgi, Gary Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25, Dublin, Ireland. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020. CycleGT: Unsupervised graph-to-text and text-to-graph generation via cycle training. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 77–88, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhi Hong, Aswathy Ajith, James Pauloski, Eamon Duede, Kyle Chard, and Ian Foster. 2023. The diminishing returns of masked language models to science. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1270–1283, Toronto, Canada. Association for Computational Linguistics.

Andrea Iovine, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2022. Cyclener: An unsupervised training approach for named entity recognition. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2916–2924, New York, NY, USA. Association for Computing Machinery.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *Poceedings of 5th International Conference on Learning Representations*.

Yuna Jeong and Eunhui Kim. 2022. Scideberta: Learning deberta for science technology documents and fine-tuning information extraction tasks. *IEEE Access*, 10:60805–60813.

Bin Ji, Shasha Li, Shaoduo Gan, Jie Yu, Jun Ma, Huijun Liu, and Jing Yang. 2022. Few-shot named entity recognition with entity-level prototypical network

enhanced by dispersedly distributed prototypes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1842–1854, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sun Kim, Rezarta Islamaj Dogan, Andrew Chatr-Aryamontri, Mike Tyers, W John Wilbur, and Donald C Comeau. 2015. Overview of biocreative v bioc track. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, Sevilla, Spain*, pages 1–9.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6248–6260, Online. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *the Sixth International Conference on Learning Representations*.

Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.

Dong-Ho Lee, Akshen Kadakia, Kangmin Tan, Mahak Agarwal, Xinyu Feng, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, and Xiang Ren. 2022. Good examples make a faster learner: Simple demonstration-based learning for low-resource NER. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2687–2700, Dublin, Ireland. Association for Computational Linguistics.

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.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2022. Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4245–4256.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023a. CodeIE: Large code generation models are better few-shot information extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.

Yueling Li, Sebastian Martschat, and Simone Paolo Ponzetto. 2023b. Multi-source (pre-)training for cross-domain measurement, unit and context extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 1–25, Toronto, Canada. Association for Computational Linguistics.

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. *Computation and Language Repository*, arXiv:1907.11692.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ruotian Ma, Zhang Lin, Xuanting Chen, Xin Zhou, Junzhe Wang, Tao Gui, Qi Zhang, Xiang Gao, and Yun Wen Chen. 2023. Coarse-to-fine few-shot learning for named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4115–4129, Toronto, Canada. Association for Computational Linguistics.

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022. Decomposed meta-learning for few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1584–1596, Dublin, Ireland. Association for Computational Linguistics.

Tasnim Mohiuddin and Shafiq Joty. 2019. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3857–3867, Minneapolis, Minnesota. Association for Computational Linguistics.

Ngoc Dang Nguyen, Lan Du, Wray Buntine, Changyou Chen, and Richard Beare. 2022. Hardness-guided domain adaptation to recognise biomedical named entities under low-resource scenarios. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4063–4071, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Venkata Prabhakara Sarath Nookala, Gaurav Verma, Subhabrata Mukherjee, and Srijan Kumar. 2023. Adversarial robustness of prompt-based few-shot learning for natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2196–2208, Toronto, Canada. Association for Computational Linguistics.

Jaehoon Oh, Sungnyun Kim, Namgyu Ho, Jin-Hwa Kim, Hwanjun Song, and Se-Young Yun. 2022. Understanding cross-domain few-shot learning based on domain similarity and few-shot difficulty. In *Advances in Neural Information Processing Systems*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *Machine Learning Repository*, arXiv:1807.03748.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. In-BoXBART: Get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States. Association for Computational Linguistics.

Yongliang Shen, Xinyin Ma, Yechun Tang, and Weiming Lu. 2021. A trigger-sense memory flow framework for joint entity and relation extraction. In *Proceedings of the Web Conference 2021*, WWW '21, page 1704–1715, New York, NY, USA. Association for Computing Machinery.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. 2019. Dual supervised learning for natural language understanding and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5472–5477, Florence, Italy. Association for Computational Linguistics.

Shang-Yu Su, Chao-Wei Huang, and Yun-Nung Chen. 2020. Towards unsupervised language understanding and generation by joint dual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 671–680, Online. Association for Computational Linguistics.

C. Sun, W. Li, J. Xiao, N. Parulian, C. Zhai, and H. Ji. 2021. Fine-grained chemical entity typing with multimodal knowledge representation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1984–1991, Los Alamitos, CA, USA. IEEE Computer Society.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Bo-Hsiang Tseng, Jianpeng Cheng, Yimai Fang, and David Vandyke. 2020. A generative model for joint natural language understanding and generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1795–1807, Online. Association for Computational Linguistics.

Richard Van Noorden. 2014. Global scientific output doubles every nine years. *Nature news blog*.

Vijay Viswanathan, Graham Neubig, and Pengfei Liu. 2021. CitationIE: Leveraging the citation graph for scientific information extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 719–731, Online. Association for Computational Linguistics.

Qingyun Wang, Manling Li, Hou Pong Chan, Lifu Huang, Julia Hockenmaier, Girish Chowdhary, and Heng Ji. 2023a. Multimedia generative script learning for task planning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 986–1008, Toronto, Canada. Association for Computational Linguistics.

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021a. ChemNER: Fine-grained chemistry named entity recognition

with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaqing Wang, Haoda Chu, Chao Zhang, and Jing Gao. 2021b. Learning from language description: Low-shot named entity recognition via decomposed framework. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1618–1630, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2021c. Meta self-training for few-shot neural sequence labeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1737–1747, New York, NY, USA. Association for Computing Machinery.

Zhuoer Wang, Marcus Collins, Nikhita Vedula, Simone Filice, Shervin Malmasi, and Oleg Rokhlenko. 2023b. Faithful low-resource data-to-text generation through cycle training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2847–2867, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weijia Xu, Xing Niu, and Marine Carpuat. 2020. Dual reconstruction: a unifying objective for semi-supervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2006–2020, Online. Association for Computational Linguistics.

Yuanyuan Xu, Zeng Yang, Linhai Zhang, Deyu Zhou, Tiandeng Wu, and Rong Zhou. 2023. Focusing, bridging and prompting for few-shot nested named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2621–2637, Toronto, Canada. Association for Computational Linguistics.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.

Li Yang, Qifan Wang, Jingang Wang, Xiaojun Quan, Fuli Feng, Yu Chen, Madian Khabsa, Sinong Wang, Zenglin Xu, and Dongfang Liu. 2023. MixPAVE: Mix-prompt tuning for few-shot product attribute value extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9978–9991, Toronto, Canada. Association for Computational Linguistics.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Hai Ye, Wenjie Li, and Lu Wang. 2019. Jointly learning semantic parser and natural language generator via dual information maximization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2090–2101, Florence, Italy. Association for Computational Linguistics.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6261–6270, Online. Association for Computational Linguistics.

A Training and Evaluation Details

	Avg. runtime	# of Parameters
RoBERTa	16min	125M
PubMedBERT	18min	109M
ScholarBERT	19min	355M
InBoXBART	58min	139M
+Valid	56min	279M
+Valid+CL	59min	279M

Table 11: Runtimne (exclude CrossNER) and Number of Model Parameters

Our baselines and model are based on the Huggingface framework (Wolf et al., 2020)[7]. Our models are trained on a single NVIDIA A100 GPU.

[7] https://github.com/huggingface/transformers

All hyperparameter settings are listed below. We optimize all models by AdamW (Loshchilov and Hutter, 2019). The runtime and number of parameters is listed in Table 11.

RoBERTa. We train a *RoBERTa-base* model with 100 epochs and a batch size 32. The learning rate is 2×10^{-5} with $\epsilon = 1 \times 10^{-6}$. We use a linear scheduler for the optimizer.

PubMedBERT. The PubMedBERT has the same model architecture as *BERT-base* with 12 transformer layers. The original checkpoint is pretrained on PubMed abstracts and full-text articles. We train a *microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext* model with 100 epochs and a batch size 32. The learning rate is 2×10^{-5} with $\epsilon = 1 \times 10^{-6}$. We use a linear scheduler for the optimizer.

ScholarBERT. The ScholarBERT is based on the same architecture as *BERT-large*. The original checkpoint is pretrained on 5,496,055 articles from 178,928 journals. The pretraining corpus has 45.3% articles about biomedicine and life sciences. We train a *globuslabs/ScholarBERT* model with 100 epochs and a batch size 32. The learning rate is 2×10^{-5} with $\epsilon = 1 \times 10^{-6}$. We use a linear scheduler for the optimizer.

InBoXBART. The InBoXBART is an instructional-tuning language model for 32 biomedical NLP tasks based on *BART-base*. We train the *cogint/in-boxbart* model with 100 epochs and a batch size 16. The learning rate is 10^{-5} with $\epsilon = 1 \times 10^{-6}$. During decoding, we use beam-search to generate results with a beam size 5. We use cosine annealing warm restarts schedule (Loshchilov and Hutter, 2017) for the optimizer.

InBoXBART+Valid. We first pretrain the self-validation model, which is based on *cogint/in-boxbart*, on the training set. The learning rate for the self-validation module is 1×10^{-5} with $\epsilon = 1 \times 10^{-6}$. We use BLUE and ROUGE to select the best model. We then train the entity extraction model and the self-validation model jointly with cross-entropy $\mathcal{L}_{\text{gen}}$ loss and reconstruction loss $\mathcal{L}_{\text{recon}}$. The final loss is $\mathcal{L} = \mathcal{L}_{\text{gen}} + 5 \cdot \mathcal{L}_{\text{recon}}$. The learning rate is 5×10^{-5} with $\epsilon = 1 \times 10^{-6}$. During decoding, we use beam-search to generate results with a beam size 5. We use cosine annealing warm restarts schedule (Loshchilov and Hutter, 2017) for the optimizer.

InBoXBART+Valid+CL. The final model is similar to *InBoXBART+Valid*. We retain the self-validation module and add a new decoder contrastive loss. The final loss is $\mathcal{L} = \mathcal{L}_{\text{gen}} + 0.2 \cdot \mathcal{L}_{\text{cl}} + 5 \cdot \mathcal{L}_{\text{recon}}$. We randomly choose 5 negative samples for each instance. The learning rate is 5×10^{-5} with $\epsilon = 1 \times 10^{-6}$. During decoding, we use beam-search to generate results with a beam size 5. We use cosine annealing warm restarts schedule (Loshchilov and Hutter, 2017) for the optimizer.

AMR-based Mention Extraction. We use AMR-parser (Fernandez Astudillo et al., 2020) to extract mentions. We treat all text spans that are linkable to Wikipedia as mentions.

NNShot and StructShot. We use the implementation from Ding et al. (2021) and choose *RoBERTa-base* as the language model.

Evaluation Metrics. We use entity-level micro-F1 for all experiments. We use the library from nereval `https://github.com/jantrienes/nereval`.

B Dataset Details

We list the entity types of ChemNER+ and CHEMET below:

- ChemNER+: Transition metals, Organic acids, Heterocyclic compounds, Organometallic compounds, Reagents for organic chemistry, Inorganic compounds, Thermodynamic properties, Aromatic compounds, Metal halides, Organic reactions, Alkylating agents, Organic compounds, Coupling reactions, Functional groups, Inorganic silicon compounds, Stereochemistry, Organohalides, Chemical properties, Catalysts, Free radicals, Alkaloids, Coordination chemistry, Ligands, Organophosphorus compounds, Reactive intermediates, Substitution reactions, Inorganic carbon compounds, Organonitrogen compounds, Biomolecules, Coordination compounds, Halogens, Chemical elements, Chlorides, Elimination reactions, Organic redox reactions, Inorganic phosphorus compounds, Organic polymers, Macrocycles, Cyclopentadienyl complexes, Substituents, Name reactions, Spiro compounds, Chemical kinetics, Organometallic chemistry, Catalysis, Organosulfur compounds, Ring forming reactions, Noble gases, Protecting

groups, Addition reactions, Carbenes, Inorganic nitrogen compounds, Non-coordinating anions, Polymerization reactions, Carbon-carbon bond forming reactions, Isomerism, Enzymes, Oxoacids, Hydrogenation catalysts

- CHEMET: Acyl Groups, Alkanes, Alkenes, Alkynes, Amides, Amines, Aryl Groups, Carbenes, Carboxylic Acids, Esters, Ethers, Heterocyclic Compounds, Ketones, Nitriles, Nitro Compounds, Organic Polymers, Organohalides, Organometallic Compounds, Other Aromatic Compounds, Other Hydrocarbons, Other Organic Acids, Other Organic Compounds, Other Organonitrogen Compounds, Other Organophosphorus Compounds, Phosphinic Acids And Derivatives, Phosphonic Acids, Phosphonic Acids And Derivatives, Polycyclic Organic Compounds, Sulfonic Acids, Thiols

The frequency for each type in the training data of both ChemNER+ and CHEMET are listed below:

- ChemNER+: Organic compounds: 183, Coupling reactions: 171, Aromatic compounds: 136, Functional groups: 120, Heterocyclic compounds: 106, Catalysts: 70, Biomolecules: 66, Chemical elements: 64, Organohalides: 63, Transition metals: 56, Chemical properties: 55, Ligands: 55, Organic acids: 48, Thermodynamic properties: 43, Inorganic compounds: 43, Coordination compounds: 37, Stereochemistry: 33, Organometallic compounds: 33, Reagents for organic chemistry: 28, Coordination chemistry: 27, Organonitrogen compounds: 26, Organic reactions: 23, Organic polymers: 23, Substitution reactions: 21, Catalysis: 20, Organic redox reactions: 18, Reactive intermediates: 13, Substituents: 13, Halogens: 12, Addition reactions: 8, Chlorides: 6, Ring forming reactions: 6, Inorganic carbon compounds: 6, Enzymes: 6, Alkaloids: 4, Organophosphorus compounds: 4, Organosulfur compounds: 4, Oxoacids: 4, Elimination reactions: 3, Carbenes: 3, Inorganic phosphorus compounds: 2, Chemical kinetics: 2, Macrocycles: 2, Noble gases: 2, Organometallic chemistry: 2, Hydrogenation catalysts: 2, Metal halides: 1, Cyclopentadienyl complexes: 1, Inorganic nitrogen compounds: 1, Protecting groups: 1, Alkylating agents: 1, Polymerization reactions: 1

- CHEMET: Other Organic Compounds: 1705, Ethers: 934, Other Aromatic Compounds: 882, Heterocyclic Compounds: 792, Alkanes: 528, Amides: 516, Other Organonitrogen Compounds: 501, Organometallic Compounds: 495, Esters: 440, Amines: 431, Ketones: 406, Polycyclic Organic Compounds: 375, Aryl Groups: 363, Organohalides: 312, Alkynes: 281, Alkenes: 266, Organic Polymers: 255, Other Hydrocarbons: 236, Other Organic Acids: 194, Other Organophosphorus Compounds: 97, Acyl Groups: 78, Nitriles: 77, Carboxylic Acids: 62, Sulfonic Acids: 37, Nitro Compounds: 26, Carbenes: 9, Phosphonic Acids And Derivatives: 4, Thiols: 2

We consider the following types as long-tail entity types for ChemNER+ and CHEMET. We list both the entity type and its frequency:

- ChemNER+: Reactive intermediates: 13, Substituents: 13, Halogens: 12, Addition reactions: 8, Chlorides: 6, Ring forming reactions: 6, Inorganic carbon compounds: 6, Enzymes: 6, Alkaloids: 4, Organophosphorus compounds: 4, Organosulfur compounds: 4, Oxoacids: 4, Elimination reactions: 3, Carbenes: 3, Inorganic phosphorus compounds: 2, Chemical kinetics: 2, Macrocycles: 2, Noble gases: 2, Organometallic chemistry: 2, Hydrogenation catalysts: 2, Metal halides: 1, Cyclopentadienyl complexes: 1, Inorganic nitrogen compounds: 1, Protecting groups: 1, Alkylating agents: 1, Polymerization reactions: 1

- CHEMET: Alkynes: 281, Alkenes: 266, Organic Polymers: 255, Other Hydrocarbons: 236, Other Organic Acids: 194, Other Organophosphorus Compounds: 97, Acyl Groups: 78, Nitriles: 77, Carboxylic Acids: 62, Sulfonic Acids: 37, Nitro Compounds: 26, Carbenes: 9, Phosphonic Acids And Derivatives: 4, Thiols: 2

C Evaluation on Whole Dataset

We conduct fully supervised training on all training sets. The results are listed in Table 12 and 13. We observe that the self-validation module

Model	Precision	Recall	F1
In-BoXBART	55.73	43.28	48.72
+ Valid	**57.49**	45.77	50.97
+ Valid + CL	57.41	**46.20**	**51.10**

Table 12: micro-F1 for ChemNER+ with the whole training set.

still improves the performance of the original In-BoXBART for two datasets. We observe that the decoder contrastive loss further improves the model performance on ChemNER+. However, adding the entity decoder contrastive loss slightly decreases it. Because there are 6561 sentences in the CHEMET dataset, which is larger than the ChemNER+ dataset, the model with the self-validation module already performs very well. Additionally, since the CHEMET model contains fewer entities per sentence than the ChemNER+ dataset and these entities are all organic compounds separated away from each other, the entity decoder contrastive loss might introduce noise into the generation results, consequently decreasing the performance.

Model	Precision	Recall	F1
In-BoXBART	64.94	41.62	50.73
+ Valid	**70.09**	**42.16**	**52.65**
+ Valid + CL	68.50	41.31	51.15

Table 13: micro-F1 for CHEMET with the whole training set.

D Scientific Artifacts

We list the licenses of the scientific artifacts used in this paper: PMC Open Access Subset (Gamble, 2017)[8] (CC BY-NC, CC BY-NC-SA, CC BY-NC-ND licenses), Huggingface Transformers (Apache License 2.0), ChemNER (no license), CHEMET[9] (MIT license), RoBERTa (cc-by-4.0), PubMedBERT (MIT license), ScholarBERT (apache-2.0), BLEU[10], ROUGE[11], InBoXBART (MIT license), brat (MIT license), and nereval (MIT license). Our usage of existing artifacts is consistent with their intended use.

[8] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[9] https://github.com/chenkaisun/MMLI1

[10] https://github.com/cocodataset/cocoapi/blob/master/license.txt

[11] https://github.com/cocodataset/cocoapi/blob/master/license.txt

E Human Annotation

The instructions for human annotations can be found in the supplementary material. Human annotators are required to annotate the chemical compound entities mentioned either in natural language or chemical formulas and other chemical related terms including reactions, catalysts, etc. We recruit two senior Ph.D. students from the Chemistry department in our university to perform human annotations. We use brat (Stenetorp et al., 2012) for all human annotations.

F Ethical Consideration

The Chem-FINESE model and corresponding models we have designed in this paper are limited to the chemical domain, and might not be applicable to other scenarios.

F.1 Usage Requirement

Our Chem-FINESE system provides investigative leads for few-shot fine-grained entity extraction for the chemical domain. Therefore, the final results are not meant to be used without any human review. However, domain experts might be able to use this tool as a research assistant in scientific discovery. In addition, our system does not perform fact-checking or incorporate any external knowledge, which remains as future work. Our model is trained on PubMed papers written in English, which might present language barriers for readers who have been historically underrepresented in the NLP/Chemical domain.

F.2 Data Collection

Our ChemNER+ sentences are based on papers from PMC Open Access Subset. Our annotation is approved by the IRB at our university. All annotators involved in the human evaluation are voluntary participants and receive a fair wage. Our dataset can only be used for non-commercial purposes based on PMC Open Access Terms of Use.