Supplementary Information

S1. SciOL Corpus

S1.1. Statistics

dataset/training corpus	tokens
Arxiv abstracts	318M
SCIBERT (Semantic Scholar)	3.17B
PubMed abstracts	3.2B
PubMed	16.8B
SciOL	14.9B

Table S1: Size comparison of scientific datasets and scientific corpora used for pretraining. (Numbers for pubmed and pubmed abstracts from [20].)

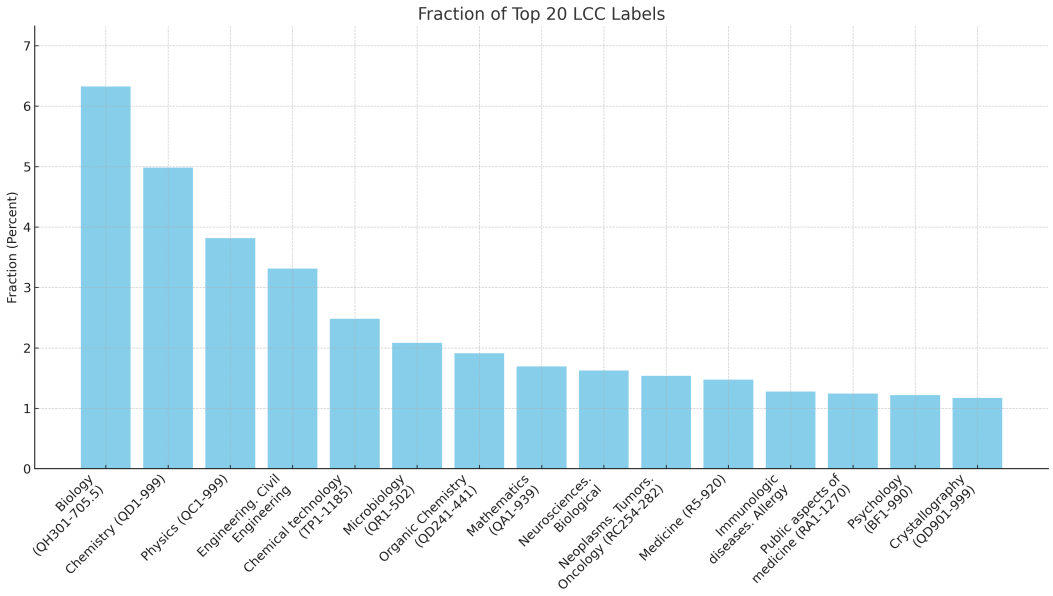


Figure S1: Top 20 labels associated with publications from SciOL in the Library of Congress Classification (LCC) system.

	PubMed	DOAJ	SciOL
# publications	0.65M	2.1M	2.75M
# text tokens	3.13B	11.73B	14.86B
# figures	2.25M	15.88M	18.13M

Table S2: Corpus statistics for the SciOL dataset and the distribution between PubMed and DOAJ. Publications are filtered by their DOI to prevent multiple occurrences.

S1.2. Quality Analysis

	WER	Substitutions	Insertions	Deletions
Our	**36.1**	7.1	13.2	**15.7**
Baseline	64.1	**2.0**	**3.4**	58.6

Table S3: Evaluation of the caption extraction quality in terms of substitutions (S), insertions (I), deletions (D) and word error rate (WER) in %. We use pdffigures2.0 [9] as baseline.

S2. SciOL Schemas

S2.1. Schema for Metadata

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "type": "object",
  "properties": {
    "doi": {
      "type": "string"
    },
    "article_url": {
      "type": "string",
      "format": "uri"
    },
    "pdf_url": {
      "type": "string",
      "format": "uri"
    },
    "license": {
      "type": "string",
      "format": "uri"
    },
    "title": {
      "type": "string"
    },
    "authors": {
      "type": "array",
      "items": {
        "type": "string"
      },
    },
    "keywords": {
      "type": "array",
      "items": {
        "type": "string"
      },
    },
    "doaj_id": {
      "type": "string"
    },
    "pmcid_id": {
      "type": "string"
    },
    "publisher": {
      "type": "string"
    },
    "issn": {
      "type": "string"
    },
    "eissn": {
      "type": "string"
    }
  },
  "required": [
    "doi",
    "pdf_url",
    "license",
    "keywords",
    "title",
    "authors"
  ]
}
```

Listing 1: JSON schema for SciOL metadata. We use a flat schema for simplicity. Depending of the articel index source (DOAJ vs. PMC), we provide different information alongside each entry in addition to the common keys defined under *required*.

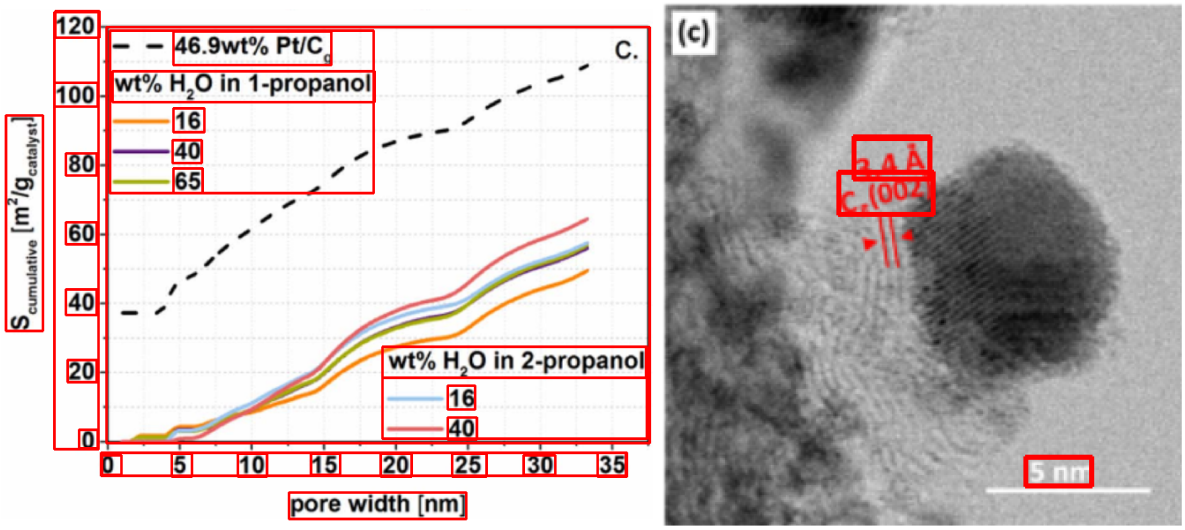
S2.2. Schema for Text Data

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "type": "object",
  "properties": {
    "doi": {
      "type": "string"
    },
    "keywords": {
      "type": "array",
      "items": {
        "type": "string"
      }
    },
    "license": {
      "type": "string"
    },
    "article": {
      "type": "object",
      "properties": {
        "title": {
          "type": "string"
        },
        "authors": {
          "type": "array",
          "items": {
            "type": "string"
          }
        },
        "abstract": {
          "type": "string"
        },
        "body_text": {
          "type": "string"
        },
        "bibliography": {
          "type": "array",
          "items": {
            "type": "string"
          }
        }
      },
      "required": [
        "title",
        "authors",
        "abstract",
        "body_text",
        "bibliography"
    ]
    }
  },
  "required": [
    "doi",
    "keywords",
    "license",
    "article"]
}
```

Listing 2: JSON schema for the textual data extracted from the PDF files.

S3. MuLMS-Img Corpus

S3.1. Example Images



Pore area distribution dependency on pore width.

1. Bright field stem of thermally treated electrospun carbon nanofibers.
2. Bright field stem for the recognition of graphitic ordered structure on the (002) plane.

Figure S2: Additional images from the MuLMS-Img corpus with retrieval queries. Text annotations are indicated by bounding boxes. Extracted from [46] (left) and [22] (right).

S3.2. Annotation Guidelines

S3.2.1 Data Collection and Splitting

MuLMS-Img was constructed from 50 publications from seven sub-domains of material science (see Table S4). Since publications contain visually similar figures, we split the corpus into a train, development and test set on publication level to prevent data leakages between the training and evaluation data.

Figures and captions are manually extracted and matched. Basic guidelines for the annotations are explained in the following.

Category	# of examples	Count Perc.
Polymers	11	22
Semiconductors	11	22
Electrolysis	4	8
Graphene	2	4
Steel	10	20
PEMFC	8	16
SOFC	4	8

Table S4: Taxonomy and domain distribution of the publications used to construct MuLMS-Img.

S3.2.2 Figure Type Classification

We define 11 fine grained classes for based on the taxonomy of UBPMC [11]. Because of the comparably few samples we join horizontal and vertical classes, e.g., horizontal bar charts and vertical bar charts. Our taxonomy is defined as follows:

- **Area Chart**: line charts, where the area between lines or the axis is emphasized, e.g., through color.
- **Bar Chart**
- **Line Chart**
- **Scatter Plot**
- **Scatter-Line Plot**: line charts, which contain additional data points, e.g. for measurements or regression. Line charts with intervals and additional points highlighted on the line should also be considered as scatter-line plots.
- **Surface Plot**: 3D plots of a topology.
- **Heatmap**: used to visualize a two dimensional area or function such as 2D projections of a topology.
- **Interval Chart**
- **Illustration**
- **Photograph/Micrograph**
- **Other**: all figures that can not be associated to one of the other classes.

S3.2.3 Optical Character Recognition (OCR) and Role Labeling

We annotate figure elements with bounding boxes and class label annotation depending on the role of the element within the figure. In addition, we transcribe the text within a bounding box, which could be used for optical character recognition. The taxonomy is defined as follows:

- **Title**
- **X-axis label**
- **Y-axis label**
- **Legend**
- **X-axis tick**.
- **Y-axis tick**
- **Label**: textual label, e.g., for additional information.
- **Plotarea**: area withing a chart where the data is visualized.
- **Text** the transcibed text using typesetting based on the LaTEX syntax.

S3.2.4 Figure Retrieval

Our aim is to create real-world textual queries that might be used in a search engine. The writing style typically deviates from the descriptive and wordy nature of captions. We ask our expert annotators to describe the figure in one or at most two consecutive sentences and be as concise as possible without getting to vague or ambiguous. There can be multiple queries for a single image.

S3.3. Corpus statistics

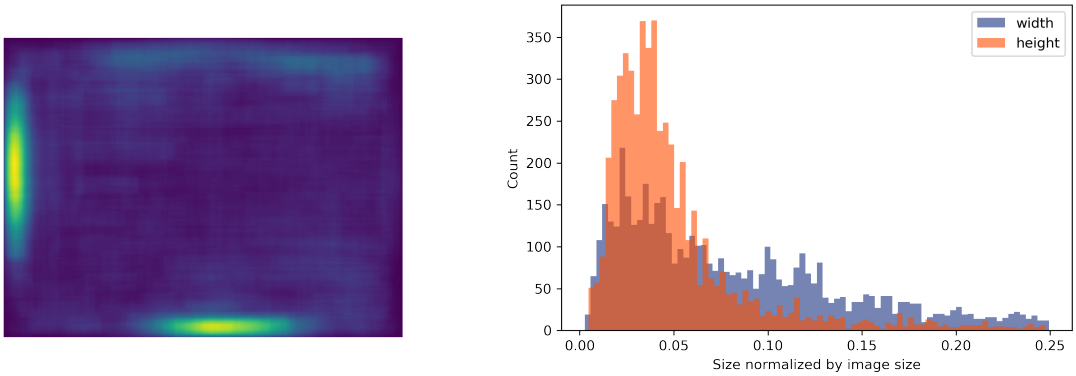


Figure S3: Location distribution of bounding boxes (left) and corresponding width and height distribution (right). Scientific figures show a concentration of text along the left and bottom image border, likely due to the position of axis labels.

	Overall	Train	Dev	Test
Total	1074	729	184	161
Line	322	222	50	50
Photography	270	189	24	57
Heatmap	22	16	2	4
Area	33	22	10	1
Surface	9	9	0	0
Scatter-line	130	58	69	3
Scatter	67	52	6	9
Illustration	156	119	17	20
Interval	29	20	5	4
Bar	32	19	1	12
Other	4	3	0	1

Table S5: Class distribution of figure type annotations in MuLMS-Img.

S4. Loss Functions

S4.1. Contrastive Matching Loss

The contrastive matching loss $\mathcal{L}_{\text{Con}}$ is given as:

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Bigg[\underbrace{\left(\sum_{i=1}^{N}\log\frac{\exp\left(\mathbf{x}_i^\top \mathbf{y}_i/\sigma\right)}{\sum_{j=1}^{M}\exp\left(\mathbf{x}_i^\top \mathbf{y}_{m,j}/\sigma\right)}\right)}_{\text{image-to-text}} + \underbrace{\left(\sum_{i=1}^{N}\log\frac{\exp\left(\mathbf{y}_i^\top \mathbf{x}_i/\sigma\right)}{\sum_{j=1}^{M}\exp\left(\mathbf{y}_i^\top \mathbf{x}_{m,j}/\sigma\right)}\right)}_{\text{text-to-image}}\Bigg]$$

where x and y are the normalized image and text representations and the subscript m indicates samples from the momentum encoders.

S4.2. Captioning Loss

The captioning loss $\mathcal{L}_{\text{Cap}}$ is given as the negative log-likelihood of the predicted token y_t conditioned on the previous token sequence $y_{<t}$ and the image latent representation x:

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T}\log P_\theta(y_t|y_{<t}, x)$$

S4.3. Joint Loss function

As joint optimization objective we optimize the equally weighted captioning and contrastive matching loss, given as:

$$\mathcal{L}_{\text{CoCa}} = \mathcal{L}_{\text{Con}} + \mathcal{L}_{\text{Cap}}$$

S5. Hyperparameter settings

S5.1. SciOL-CoCa

Hyperparameter	
Optimizer	AdamW
LR decay schedule	cosine
Train steps	300k
Batch size	1600
Learning rate	5e-4
Warmup-steps	5000
Weight decay	0.2
Input size	300
Augmentation	RandAugment(1,1) [10], Image patch masking (50%)

Table S6: Hyperparameters for pre-training CoCa on SciOL for retrieval and captioning.

S5.2. Optical Character Recognition

Hyperparameter	
Optimizer	AdamW
LR decay schedule	cosine
Epochs	10
Train batch size	64
Learning rate	5e-5
Weight decay	1e-2

Table S7: Hyperparameters for finetuning TrOCR on MuLMS-Img and UBPMC for text recognition.

Hyperparameter	MuLMS-Img	UBPMC
Optimizer	SGD	SGD
LR decay schedule	cosine	cosine
Train steps	7500	2000
Train batch size	32	32
Learning rate	0.01	0.01
Warmup-steps	500	200
Weight decay	1e-4	1e-4
input size	1200	1200

Table S8: Hyperparameters for training faster-RCNN on MuLMS-Img and UBPMC for text detection.

S6. Experiments

S6.1. Finetuning without Patchdropout

	text-to-image			image-to-text		
	R@1	R@5	R@10	R@1	R@5	R@10
SciOL-CoCa	9.8	18.1	22.3	10.2	19.2	23.9
SciOL-CoCa + no patch-dropout	10.0	19.1	23.5	10.3	19.9	24.6

Table S9: Influence of final tuning episode without patch dropout measured by zero shot scientific figure retrieval on the SciOL test set (using captions as queries).

S6.2. Optical Character Recognition

	MuLMS-Img		UBPMC	
	AP	AR	AP	AR
Faster R-CNN	79.8	83.8	62.0	70.6
+ CoordConv	80.2	84.0	59.4	68.3
+ text anchors	81.4	85.4	62.6	70.8
+ augmentation	81.6	85.5	62.9	70.5

Table S10: Evaluation results for text detection on the MuLMS-Img and UBPMC text bounding box annotations exclusively trained on MuLMS-Img or UBPMC.

S6.3. Figure Type Classification

	MuLMS-Img				UBPMC			
	F1-Macro	Precision	Recall	Micro scores	F1-Macro	Precision	Recall	Micro scores
Biomed-CLIP	17.7	21.6	23.0	26.1	32.1	33.7	52.2	34.7
CoCa-Vit-B32	33.9	36.0	41.3	54.6	36.5	41.9	40.5	44.0
Sciol-CoCa	39.8	49.3	43.7	61.4	39.4	37.8	54.1	46.0

Table S11: Detailed performance scores for figure type classification on MuLMS-Img and UBPMC.

S6.4. Captioning on HCI alt-text

	HCI-alt-text charts					
	Rouge			BERTScore		
	P	R	F1	P	R	F1
CoCa	16.9	6.0	7.7	77.5	79.1	78.2
SciOL-CoCa	26.8	12.5	14.4	81.8	80.9	81.3

Table S12: Zero shot figure captioning on the HCI-alt-text.