DEPARTMENT OF ENGINEERING MATHEMATICS

Precision-First Information Extraction of Polymer Properties from Scientific Literature

Anshul Gupta

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Friday 5th September, 2025

Supervisor: Dr Sébastien Rochat

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Anshul Gupta, Friday 5th September, 2025

Contents

1 Introduction **1**

1.1 Context and Motivation 1

1.2 Problem Statement 1

1.3 Research Objectives and Questions 1

2 Background **3**

2.1 Information extraction in materials science 3

2.2 Evolution of Polymer Corpora and Modelling Approaches 3

2.3 Motivation for Weak Supervision and Contribution of this Study 4

3 Methodology **6**

3.1 Corpus construction 6

3.1.1 Sources and scope 6

3.1.2 Annotation Schema 7

3.1.3 Data Harvesting and Gazetteer Construction 8

3.1.4 Automated Annotation 10

3.1.5 Preprocessing 11

3.2 Conditional Random Fields 12

3.3 Bidirectional LSTM and Transformer-based Models 13

3.3.1 BiLSTM baseline (Word only) 13

3.3.2 Character-aware BiLSTMs (CharCNN vs CharBiLSTM) 14

3.3.3 BiLSTM + CRF (CharCNN) 15

3.3.4 SciBERT-based models 15

4 Results and Discussion **17**

4.1 Overall Performance 17

4.1.1 Evaluation Framework 17

4.1.2 Corpus-level comparison 17

4.1.3 Model family trends 18

4.1.4 Per-label analysis 18

4.1.5 Interpretation and transition 18

4.2 Model-wise Comparison 19

4.2.1 Rationale for model-level evaluation 19

4.2.2 Comparative performance across models 19

4.2.3 Interpretation 19

4.3 Error Analysis 21
4.3.1 Distribution of error types across entities 21
4.3.2 Sentence-level error buckets 21
4.3.3 Interpretation 23
4.4 Qualitative Examples 23
4.4.1 Bulk corpus 24
4.4.2 Gold corpus 24
4.4.3 Ramprasad corpus 25
4.4.4 Cross-model progression examples 25
4.4.5 Interpretation 26
4.5 Results Summary 26

5 Conclusions and Future Work **28**
5.1 Revisiting Research Questions and Aims 28
5.2 Key Findings and Contributions 29
5.3 Limitations 29
5.4 Future Work 30

References **31**

List of Figures

3.1 Entity Coverage (token-level, raw) Across Corpora 6

3.2 End-to-end pipeline for corpus construction 12

4.1 Heatmap of per-entity F1 scores for best-performing models on Bulk (BiLSTM+CRF), Gold (BiLSTM+CRF), and Ramprasad (SciBERT fine-tuned) 18

4.2 Weighted and macro F1 by model on Bulk corpus 20

4.3 Weighted and macro F1 by model on Gold corpus 20

4.4 Weighted and macro F1 by model on Ramprasad corpus 20

4.5 Distribution of error types by entity on Bulk corpus 22

4.6 Distribution of error types by entity on Gold corpus 22

4.7 Distribution of error types by entity on Ramprasad corpus 22

4.8 Sentence-level prediction buckets for best-performing models on Bulk (BiLSTM+CRF), Gold (BiLSTM+CRF), and Ramprasad (SciBERT fine-tuned) 23

List of Tables

3.1 Overview of corpora 6

3.2 Relation schema used in annotation. 7

3.3 Example annotation of a sentence fragment. 7

3.4 Coverage statistics of harvested gazetteers after cleaning and regex generation 9

3.5 Example refinement statistics aggregated across annotated corpora 11

3.6 Segment-level preprocessing statistics for Ramprasad, Gold, and Bulk corpora at word and piece levels. 12

3.7 Feature groups and their roles in the CRF model. 13

4.1 Best model performance per corpus (test set). 17

Abstract

The growing field of polymer informatics depends on structured, high-quality data to support discovery and modelling. However, most existing databases are either incomplete or restricted by commercial licensing. Open-access literature contains abundant descriptions of polymers and their properties, but converting these into usable datasets requires both careful annotation and robust modelling.

This project presents a complete pipeline for property extraction from scientific papers. A unified schema was designed covering polymers, properties, values, units, methods, solvents, and additives. External resources such as the Ramprasad abstract dataset were remapped into this schema, enriching entity coverage by grouping contextually similar classes (e.g., polymer families, organic and inorganic materials mapped into POLYMER or ADDITIVE). The addition of SOLVENT and METHOD entities, together with lightweight relation labels such as has_unit and describes_property, allows richer and more useful information to be extracted. Two new corpora were created: a Gold set of 43 full-text polystyrene papers, anchored in five manually verified articles and the remainder weakly labelled, and a Bulk set of over 3,000 weakly labelled full texts across diverse polymers. Weak labelling followed a precision-first strategy to minimise noise while expanding coverage.

Eight neural sequence labelling models were compared, ranging from word-level BiLSTM baselines with GloVe and domain embeddings, through character-aware hybrids, to BiLSTM-CRF and SciBERT-based models. Results show clear improvements over the CRF baseline: F1 scores reached 0.69 on the remapped abstract corpus, 0.87 on the polystyrene Gold set, and 0.97 on the Bulk corpus. Analysis indicates that vocabulary breadth and structural coverage often matter more than class balancing alone, and that full-text data provides clear gains in context richness. The work delivers constructed corpora, a reproducible annotation framework, and systematic model benchmarks, demonstrating how weak labelling, entity enrichment, and relation-aware design can support polymer property extraction at scale.

Acknowledgements

I would like to express my sincere gratitude to Dr Sébastien Rochat, my project supervisor, for invaluable guidance and constructive advice on the objectives and content of this report. I am also deeply grateful to my family for their unwavering support throughout the project. Finally, I thank the University of Bristol for providing the facilities and resources that made this work possible.

Ethics Statement

After discussion with my project supervisor, Dr Sébastien Rochat, the work analyses only publicly available literature and therefore does not require formal ethics approval.

I have completed the ethics test on Blackboard. My score is 12/12.

Chapter 1

Introduction

1.1 Context and Motivation

Polymers form the backbone of modern materials science and are used in everything from packaging and construction to advanced electronics and biomedicine. Their usefulness comes from the fact that the same chemical backbone can lead to very different behaviours depending on molecular weight, additives, processing methods, or solvent environment. For example, properties such as the glass transition temperature, refractive index, and mechanical strength are critical when selecting materials for applications in electronics, optics, or drug delivery. Designing new polymers with tailored behaviour therefore requires not only chemical intuition but also access to structured and reliable property data.

Despite this importance, much of the knowledge about polymers and their intrinsic properties remains scattered across thousands of research papers, patents, and reports. Databases that exist are often commercial and closed, while academic collections are limited in scope and coverage. This lack of open and structured data makes it difficult to apply modern data science methods, such as machine learning, which rely on large and representative datasets. Researchers who want to study structure-property relationships often must manually search and annotate literature, a process that is slow, error-prone, and cannot scale to the pace at which new articles are published.

Natural Language Processing (NLP) offers a route to automate this process by extracting entities such as polymers, experimental methods, values, and units directly from text. Recent years have seen significant progress in scientific NLP, especially with the development of domain-specific word embeddings and transformer models such as SciBERT. These advances suggest that large-scale, automated extraction of polymer property data is possible, provided that the text is carefully pre-processed, annotated, and modelled. Building such a pipeline can help to bridge the gap between unstructured literature and structured databases, thereby accelerating research in polymer informatics.

1.2 Problem Statement

The central problem addressed in this project is the absence of open, structured, and context-rich datasets that capture intrinsic polymer properties in a machine-readable format. While prior work in materials informatics has demonstrated the value of text mining for metals, alloys, and metal–organic frameworks, polymers present a unique challenge. Polymer names are highly variable, often represented with prefixes, brackets, or shorthand notations, and property mentions are embedded in experimental descriptions rather than presented in standardised tables. As a result, existing approaches that rely on abstract-only text or narrow sets of entity types fail to capture the full context in which polymer properties are reported.

This lack of structured coverage limits both data availability and model generalisation. Without a diverse and representative training corpus, sequence labelling models struggle to recognise new polymer names, values, or contextual cues. Furthermore, many available datasets for polymer NLP, such as the Ramprasad group abstracts, contain restricted entity categories that do not fully reflect the breadth of experimental reporting. To make progress, it is necessary to construct corpora that go beyond abstracts and that capture a wider range of entity types, including solvents, methods, and additives, which directly influence reported property values.

The specific problem this thesis tackles is therefore twofold:

1. How to create and harmonise corpora that provide high-quality training data for polymer property extraction, including both manually verified and weakly labelled full-text resources.

2. How to evaluate and improve sequence labelling models such that they can generalise across polymers, properties, and reporting styles, ultimately enabling the automated construction of large, open datasets for polymer informatics.

1.3 Research Objectives and Questions

The overarching aim of this project is to design and evaluate a complete pipeline for automated extraction of polymer entities and their associated properties from open-access scientific literature. To achieve this, the work is guided by the following objectives:

1. **Corpus Construction and Enrichment**

   Develop a high-quality gold corpus centred on polystyrene, with manual verification of seed documents and weak labelling applied to the remainder.
   Build a large bulk corpus of over 3,000 weakly annotated full papers covering a diverse set of polymers.
   Remap the Ramprasad abstract dataset into a unified schema with richer entity coverage, enabling broader vocabulary and pattern diversity.

2. **Schema and Relation Design**

   Define a schema that captures both core polymer entities and supporting experimental context, including properties, values, units, methods, solvents, and additives.
   Incorporate lightweight relation labels (e.g. has_unit) to improve the usability of extracted information in downstream applications.

3. **Model Development and Benchmarking**

   Establish a feature-engineered CRF baseline as a reference point.
   Implement and compare five LSTM variants (word-level, character-aware, and CRF-integrated) and two SciBERT-based configurations.
   Evaluate performance using strict BILOU weighted- and macro-F1 metrics, with resampling to assess statistical significance.

4. **Analysis of Data and Model Behaviour**

   Assess the relative benefits of abstract-level versus full-text training data.
   Analyse the effect of schema enrichment on model generalisation.
   Investigate the impact of class imbalance handling, vocabulary breadth, and entity coverage on extraction performance.

From these objectives, the research is structured around the following guiding questions:

RQ1: How can weakly labelled full-text corpora be constructed in a way that balances coverage and precision?

RQ2: Does expanding the schema with additional entity types and relations improve the contextual richness of extracted information?

RQ3: What performance gains can be achieved by moving from feature-based CRF models to recurrent and transformer-based architectures?

RQ4: How do models trained on abstracts compare to those trained on full papers in terms of generalisation and strict F1 performance?

RQ5: Which factors such as class imbalance handling, vocabulary diversity, or schema enrichment contribute most to reliable polymer property extraction?

Chapter 2
Background

2.1 Information extraction in materials science

The extraction of structured information from scientific articles has long been recognised as a bottleneck in materials informatics. Conventional resources such as PolyInfo and Polymer Genome provide curated collections of polymer structures and property data, but they depend heavily on manual expert annotation. This reliance limits their scalability and slows their responsiveness to the rapidly expanding body of literature [1], [2]. As publication volumes in chemistry and materials science continue to increase, automated text-mining approaches have become essential to complement curated repositories and accelerate data-driven discovery.

Neighbouring domains have shown how natural language processing (NLP) can be applied to achieve large-scale information extraction. In the case of metal–organic frameworks (MOFs), Park and colleagues demonstrated that pipelines built around entity recognition and relation extraction could recover synthesis conditions and property data from heterogeneous publications [3]. Their later work extended these methods to construct knowledge graphs of synthesis, providing structured representations of experimental parameters and illustrating how text mining can support hypothesis generation [4]. These contributions confirmed that automated approaches can reliably extract structured knowledge at scale, reducing the need for manual curation.

Polymers, however, present unique challenges compared with crystalline solids or MOFs. Their nomenclature is unusually diverse, spanning systematic chemical names such as poly(ethylene terephthalate), shorthand forms such as PET, and domain-specific abbreviations such as PS for polystyrene. Property descriptions are often embedded in complex expressions, with numeric values expressed as ranges, ratios, or approximate quantities, and units reported in multiple notations. These factors complicate tokenisation, entity recognition, and boundary detection. Moreover, polymers lack universally standardised representations, meaning that contextual cues often determine whether a phrase denotes a polymer, an additive, or a property descriptor.

Because of these complexities, progress in polymer information extraction has lagged behind other areas of materials science. In contrast, research in inorganic and MOF domains has produced text-mined datasets and predictive frameworks [3], [4]. More broadly, recent work in materials science has demonstrated the potential of automated pipelines to extract entities and relations at scale [5], while studies in chemistry have explored the use of large language models such as ChatGPT to assist in text mining and synthesis prediction [6]. The EPSRC Seed Corn report on polymer informatics highlighted that despite these advances, data scarcity and the absence of scalable extraction methods continue to constrain the integration of polymer knowledge into discovery pipelines [7]. This recognition provided a direct motivation for the present study, which explores weakly supervised corpus construction and neural sequence models tailored to the demands of polymer literature.

2.2 Evolution of Polymer Corpora and Modelling Approaches

The development of polymer information extraction has been closely tied to the availability of annotated corpora. One of the earliest systematic datasets was introduced by Afzal et al. (2019), covering around 200 annotated abstracts with entities such as polymers, properties, and values [8]. This dataset provided a starting point for benchmarking polymer named entity recognition (NER), although its abstract-only scope restricted its generalisation to full-length papers, where property mentions are often embedded in complex descriptions of experiments.

Subsequent contributions began addressing specific challenges of polymer terminology. Shetty and Ramprasad (2021) proposed methods for entity normalisation, linking systematic polymer names, abbreviations, and trade names to canonical forms [9]. While not a corpus release, this work highlighted the complexity of polymer nomenclature, which remains a critical obstacle in domain text mining. More recently, Cheung et al. (2023) released POLYIE, a large-scale corpus of 146 full-text polymer articles

annotated with entities and relations [10]. POLYIE marked a significant step forward in scope and scale, though its relatively recent release means it has not yet become a widely adopted resource for downstream property extraction.

Parallel to corpus development, modelling approaches advanced significantly. Early systems relied on Conditional Random Fields (CRFs) [11], which leveraged handcrafted lexical, orthographic, and dictionary-based features. CRFs were reliable for predictable patterns such as numeric values and standard units but struggled with the diverse morphology of polymer names.

The introduction of distributed word embeddings enabled a new generation of neural sequence models. GloVe [12] and Mat2Vec [13] provided representations that captured semantic similarity, and when combined with bidirectional LSTMs, they improved recall and span integrity compared with CRFs. Word-only BiLSTMs nevertheless showed limitations when faced with morphological variation in polymers and units.

To address this, character-aware hybrids were developed. Architectures combining character-level CNNs or BiLSTMs with word embeddings captured internal subword patterns such as prefixes, suffixes, and systematic name fragments, thereby improving robustness to unseen tokens [14]. When paired with CRF decoders, these models further improved boundary consistency by enforcing structural legality in predicted spans [15]. This combination became a practical middle ground, balancing contextual modelling with structured decoding.

More recently, transformer-based models such as BERT [16] and SciBERT [17] have set the benchmark for scientific NER. SciBERT, trained on a multi-disciplinary scientific corpus, consistently improved recognition of long entities and abbreviations compared with recurrent models, and domain-adaptive variants such as MatSciBERT [18] and MaterialsBERT [19] have demonstrated further gains when tuned on materials-specific corpora. These models represent the current state of the art in scientific text mining, but they remain computationally intensive and less stable when trained on small or weakly supervised corpora. In this project, transformers were therefore treated not as baselines but as comparators, providing a reference point against which the practicality and precision of lighter BiLSTM–CRF architectures could be assessed.

Taken together, this trajectory shows a clear progression: from feature-driven CRFs with small corpora, through neural recurrent models that integrated semantic embeddings and character-level encoders, to contextual transformers that model sentence-wide dependencies. The present work built upon this progression by constructing two weakly labelled full-text corpora (Bulk and Gold) and systematically testing architectures across this spectrum, with the goal of determining which designs best balance scalability, precision, and robustness for polymer property extraction.

2.3 Motivation for Weak Supervision and Contribution of this Study

Despite steady progress in corpus development and modelling, the pace of advancement in polymer information extraction has been limited by the scarcity of large, high-quality annotated resources. Manual annotation, while precise, is expensive, time-consuming, and difficult to scale, particularly in a domain where entities span complex nomenclature and measurement conventions. This creates a tension between the need for broad coverage and the necessity of maintaining reliable labels.

The EPSRC Seed Corn project on polymer informatics explicitly recognised these barriers, emphasising the shortage of structured data as a primary bottleneck in accelerating polymer discovery and design [7]. Existing databases such as PolyInfo and Polymer Genome have demonstrated the value of structured polymer property repositories, but their reliance on expert-driven curation constrains their scalability [1], [2]. A sustainable alternative requires methods that can construct large corpora with minimal manual effort while still preserving sufficient quality for training statistical models.

Weak supervision offers one such solution. Instead of full manual annotation, heuristic rules, distant supervision from databases, and semi-automatic pattern analysis can generate approximate labels across large volumes of text. While these annotations may not reach the accuracy of gold-standard corpora, they enable orders of magnitude greater coverage at a fraction of the cost. When coupled with rigorous preprocessing and schema validation, weakly labelled corpora can serve as effective training grounds for

models that prioritise precision over recall. This distinction is critical in scientific information extraction, where incorrect predictions may mislead downstream analyses more than missing values.

Recent work in adjacent domains has shown that weak supervision can be made reliable when carefully constrained. Zheng et al. (2023) demonstrated that large language models can assist in chemistry-related text mining, highlighting the potential of semi-automatic annotation pipelines [6]. In materials science more broadly, text mining projects have successfully integrated heuristics with neural models to construct knowledge graphs and property databases at scale [4], [5]. These studies illustrate that weak supervision, while imperfect, can bridge the gap between data scarcity and the requirements of modern deep learning approaches.

This project was designed within that context. Two complementary corpora were constructed: Bulk, a large-scale weakly labelled collection spanning over three thousand full-text papers across diverse polymers, and Gold, a smaller but domain-focused corpus of forty-three polystyrene papers labelled with weak supervision guided by manual pattern analysis. Both were harmonised under a unified seven-entity schema and preprocessed into training-ready formats. This design enabled systematic evaluation of how scale, domain focus, and supervision quality interact in polymer property extraction.

The study's methodological contribution lay in systematically progressing from CRF baselines through BiLSTM variants to SciBERT contextual embeddings, providing a controlled comparison across modelling paradigms. In doing so, it tested whether transformer-level contextualisation offered measurable gains under weak supervision, and whether character-aware BiLSTM–CRF architectures could provide a more stable and computationally efficient alternative. By explicitly prioritising precision-first extraction, the work balanced scalability with reliability, addressing a core limitation of existing resources.

Together, these contributions extend the field by demonstrating that weak supervision can produce corpora sufficiently reliable for training high-performing models, particularly when coupled with character-aware recurrent architectures. The findings also highlight when more complex transformer-based methods add value, and when their cost and instability outweigh their benefits. In this way, the project advances both the data and modelling foundations for polymer information extraction, offering a pathway toward more scalable and precise literature mining.

Chapter 3

Methodology

3.1 Corpus construction

3.1.1 Sources and scope

All corpora used in this project are fully open access. Three datasets were constructed or adapted to support model development and evaluation under a unified schema of POLYMER, PROP_TAG, VALUE, UNIT, METHOD, SOLVENT, ADDITIVE.

Gold corpus. A collection of 43 full-text polystyrene articles, with five manually annotated to seed a weak labelling pipeline. Remaining documents were automatically labelled under a precision-first strategy to minimise noise. This corpus acts as a focused, domain-specific benchmark.

Bulk corpus. A large-scale set of 3,119 full papers spanning diverse polymers. The same weak-labelling pipeline was applied, yielding broad coverage across vocabulary and structural patterns. This corpus provides scale for training higher-capacity models.

Ramprasad abstracts. A corpus of 765 abstracts from the Ramprasad group [8], harmonised into the unified schema. Due to schema differences, SOLVENT and METHOD entities are not present in this set. It provides a contrast between abstract-only and full-text training.

Table 3.1: Overview of corpora

Corpus	Documents	Tokens (approx)	Entity spans (approx)	Annotation type
Gold corpus	43	∼0.9M	∼11.8k	5 manual + 38 weakly labelled
Bulk corpus	3,119	∼65M	∼600.8k	Weakly labelled
Ramprasad abstracts	765	∼0.18M	∼19.3k	Remapped from Ramprasad schema

The heat map (Figure 3.1) below compares entity coverage statistics across the three corpora (Ramprasad, Gold, Bulk). Rows correspond to corpora and columns summarise the number of labeled entities, providing a concise overview of dataset scale.

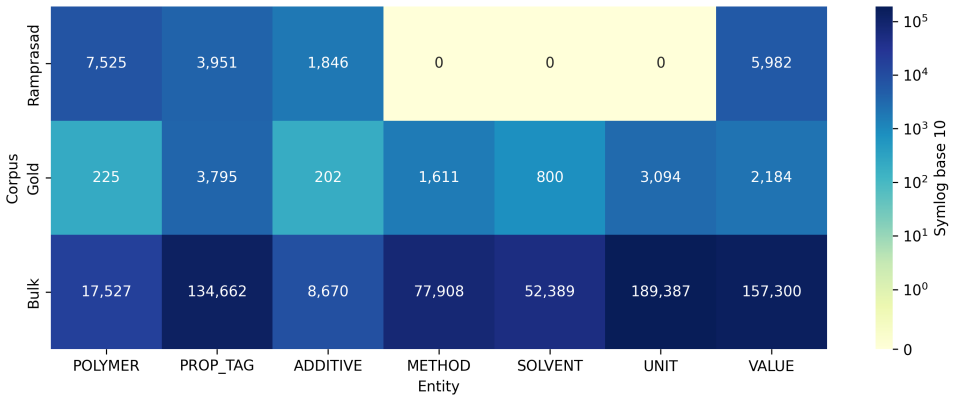

Figure 3.1: Entity Coverage (token-level, raw) Across Corpora

3.1.2 Annotation Schema

The annotation schema was designed to capture both core polymer entities and the context in which properties are reported. Seven entity types were defined, reflecting the most common units of meaning in materials science articles:

POLYMER: names of polymers or copolymers (e.g., polystyrene, PMMA).
PROP_TAG: mentions of intrinsic properties (e.g., glass transition temperature, density).
VALUE: numeric measurements or descriptors (e.g., 1.25, 210).
UNIT: associated measurement units (e.g., °C, g·mol^{-1}, MPa).
METHOD: techniques or approaches used for measurement (e.g., DSC, NMR, molecular dynamics).
SOLVENT: solvents appearing in the context of polymer experiments (e.g., chloroform, dimethylformamide, ethanol).
ADDITIVE: other materials incorporated into the polymer system (e.g., fillers, stabilisers, nanoparticles).

Alongside entities, a small but expressive set of relations was defined to connect annotations into structured triples. These relations are summarised in Table 3.2.

Table 3.2: Relation schema used in annotation.

Relation	**Head → Tail**	**Description**
has_unit	VALUE → UNIT	Links a numeric value to its measurement unit.
describes_property	VALUE → PROP_TAG	Connects a value to the property it quantifies.
about_polymer	PROP_TAG → POLYMER	Assigns a property mention to the relevant polymer system.
has_additive	POLYMER → ADDITIVE	Indicates that a polymer system includes an additive.
measured_by	VALUE → METHOD	Specifies the method or technique used to obtain a value.

This combined schema enables both entity recognition and relation extraction, supporting the transformation of raw text into a structured relational graph where polymers, properties, values, and conditions are explicitly linked. To illustrate this schema in practice, the following example is drawn from a full-text article:

"Blends of polystyrene (PS) and poly(methyl methacrylate) (PMMA) showed an increase in glass transition temperature (Tg) from 75 °C to 89 °C when measured by differential scanning calorimetry (DSC)."

The corresponding annotations are shown in Table 3.3.

Table 3.3: Example annotation of a sentence fragment.

Text span	**Entity type**	**Relation(s)**
Polystyrene (PS)	POLYMER	about_polymer (linked via PROP_TAG)
Poly(methyl methacrylate) (PMMA)	POLYMER	about_polymer (linked via PROP_TAG)
Glass transition temperature (Tg)	PROP_TAG	describes_property (from VALUE); about_polymer (to POLYMER)
75 °C	VALUE + UNIT	has_unit (75 → °C); describes_property (to PROP_TAG)
89 °C	VALUE + UNIT	has_unit (89 → °C); describes_property (to PROP_TAG)
Differential scanning calorimetry (DSC)	METHOD	measured_by (from VALUE)

This example demonstrates how the schema captures not only individual mentions but also their contextual relationships, which is essential for building structured polymer property datasets.

3.1.3 Data Harvesting and Gazetteer Construction

The annotation process was supported by a foundation of structured resources derived from open-access sources. Two key steps were required: (i) harvesting raw scientific texts to build the Gold and Bulk corpora, and (ii) constructing gazetteers to support weak labelling of entities. This section documents how full texts were gathered and how initial lexicons (gazetteers) were derived to support weak labelling.

Data Harvesting

DOI harvesting via Crossref. Candidate articles were first identified by querying the Crossref REST API (https://api.crossref.org) using journal ISSNs and bibliographic filters. Cursor-based pagination was applied to retrieve batches of DOIs matching the query. This ensured wide coverage of articles mentioning polystyrene or related polymers in the bibliographic metadata.

PMC-indexed article filtering via Europe PMC. The retrieved DOIs were then filtered through the Europe PMC search API (`https://www.ebi.ac.uk/europepmc/webservices/rest`). Only research articles indexed in PMC were retained.

Additional filters required a mention of "polystyrene" or "poly(styrene)" in the title or abstract together with at least one property cue (e.g., density, ρ, specific gravity, glass transition, Tg, refractive index, RI, n_D). A negative keyword list was applied to exclude common out-of-scope topics such as OLEDs, fluorophores, or drug delivery.

Full-text retrieval via NCBI EFetch (PMC). Full-text XML files for the retained PMCIDs were downloaded using the NCBI E-utilities EFetch endpoint (https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi) with parameters db=pmc, retmode=xml, and rettype=full. Files were stored locally, preserving PMCID identifiers for traceability.

From this process, two corpora were constructed:

- **Gold corpus**: 43 full-text articles focused on polystyrene.
- **Bulk corpus**: 3,119 full-text articles covering a wide range of polymers.

Gazetteer construction

Built four gazetteers (POLYMER, ADDITIVE, METHOD, SOLVENT) with a precision-first strategy, combining Wikipedia category harvesting, curated acronym whitelists, ontology traversal for methods, and PubChem synonym expansion for chemicals. Regex patterns were then generated with context gates for ambiguous acronyms.

APIs used

- Wikipedia API: https://en.wikipedia.org/w/api.php
- EBI OLS4 (CHMO ontology): https://www.ebi.ac.uk/ols4/api
- PubChem PUG REST: https://pubchem.ncbi.nlm.nih.gov/rest/pug

Category harvesting and title filtering (Wikipedia). Seed category sets were defined for each group (for example, *Category:Polymers; Category:Acrylic polymers; Category:Elastomers; Category:Flame retardants; Category:Plasticizers; Category:Thermal analysis; Spectroscopy; Microscopy; Category:Solvents; Organic solvents; Aromatic solvents*). From each seed, pages were collected with recursive traversal (depth up to two), followed by strict title-level filters:

- Drop namespaces and lists (e.g., Category:, Template:, "List of", User:).
- Remove chemical-formula-like strings and short all-caps (e.g., "C6H6", "TiO2", "PS" when unsafe).
- Exclude generic or non-substance pages (history, theory, industry, exposure, handbooks, etc.).
- For polymers, require shapes like poly(...) or poly-..., or explicit copolymer terms; reject generic "polymer…" phrases.

- Keep natural polymers via a small whitelist (e.g., cellulose, chitosan).

Acronyms with whitelists and context gates. Acronyms were harvested from article intros and intersected with curated safelists:

- **Polymers:** only keep from a strict whitelist (e.g., PS, PMMA, PVDF, PEEK, etc.).
- **Solvents, methods, additives:** start from curated seeds (e.g., THF, DMF, NMP; DSC, TGA, NMR; APP, DOPO, CNT) and prune fragments (Roman numerals, elemental symbols) to avoid noise.
- **Ambiguity gating:** ambiguous acronyms are gated by local context in the final regex (e.g., APP/GO/BN must co-occur with "flame retardant", "phr", "nanoparticle"; IPA/DCM must co-occur with solvent cues; BET/DMA/TMA/SEC must appear near method cues).

Methods from CHMO (OLS). For METHOD, the CHMO ontology was queried via OLS4. Seed labels (e.g., microscopy, spectroscopy, chromatography, calorimetry, thermogravimetry, scattering, diffraction, rheology, ellipsometry, porosimetry, photoelectron spectroscopy, permeation, tomography) were located, then a breadth-first traversal (up to depth four) collected child terms and synonyms. Terms were filtered with the same method-inclusion rules and noise filters. If CHMO retrieval fails, the code falls back to the Wikipedia pipeline.

PubChem synonym expansion (solvents and additives). For SOLVENT and ADDITIVE, the initial names were expanded with PubChem synonyms (up to 2 CIDs per seed and up to 200 synonyms per CID). Synonyms were aggressively cleaned to remove vendor marks, grades, amounts/units, multi-entry bundles, and generic phrases (e.g., "technical grade", "solution"). A second pass re-applied group-specific filters to keep only plausible chemical names.

Regex pattern generation. Names were converted to word-bounded, hyphen- and space-tolerant regex with escape_word_bounded_regex (appendix), then acronyms were added:

- **Polymers:** ambiguous acronyms are gated by a polymer-context window; safelist acronyms are matched plainly.
- **Solvents, methods, additives:** ambiguous acronyms are gated by their group-specific context windows.

Length caps were applied (e.g., 100 characters for long CHMO method labels), and duplicate patterns were removed case-insensitively.

Table 3.4: Coverage statistics of harvested gazetteers after cleaning and regex generation

Group	Names kept	Acronym seeds kept	Regex patterns written
Polymers	467	24	491
Additives	169	22	191
Methods	2,151	21	2,172
Solvents	285	11	296

Property and unit catalogues (manual). The property catalogue (PROP_TAG) and unit list (UNIT) were manually curated by close reading of the five manually annotated gold papers. These lists capture the canonical property terms used in polymer literature (for example, glass transition temperature, refractive index, molecular weight) and the units and notations observed (for example, °C, MPa, $g \cdot mol^{-1}$, wt%). They were then normalised and turned into compact regex patterns for the weak labeller.

All lists were de-duplicated, lower-cased for matching where appropriate, and filtered with conservative rules to avoid ambiguous acronyms unless supported by nearby context terms (for example, treat "PS" as POLYMER only when the surrounding window contains "polystyrene", "blend", "copolymer", or another clear cue). Final gazetteers were stored as plain text and regex-ready patterns for the weak labeller.

3.1.4 Automated Annotation

To scale annotation beyond the five manually labelled seed papers, a weak-labelling pipeline was developed. The design philosophy was precision-first, ensuring that automatically labelled corpora retained high reliability even at the cost of reduced recall. This decision was motivated by the downstream requirement for noise-minimised training data, where false positives are often more harmful than missing labels.

Gazetteer-driven tagging. The foundation of the annotator was the gazetteers described in Section 3.1.3. Each list was compiled into case-insensitive regex patterns, with additional rules for ambiguity resolution. Acronyms such as PS were only accepted as polymers when accompanied by contextual cues (e.g., "blend", "copolymer", "matrix"), while method acronyms like DMA were required to appear near property cues (e.g., "measured by", "tested with"). This gating strategy sharply reduced spurious matches.

Property-anchored value extraction. Numeric values were identified with a regex supporting decimals, ranges, scientific notation, and tolerance markers (±, ˜). Each candidate VALUE was paired with a UNIT token when present, with special handling for multi-token cases (e.g., "° C", "wt . % "). Units were then normalised into canonical forms ("°C" → "°C", "W/mK" → "W m^{-1} K^{-1}"). Each VALUE was linked to the nearest PROP_TAG in the same sentence window. The property catalogue, curated from the five manually annotated seed papers, acted as a control list to filter values by expected units (e.g., T_g kept only if in °C or K; molecular weight required mass-per-mole units). Implausible values (e.g., $T_g < -150$°C or density > 8 g cm^{-3}) were discarded via plausibility bands.

Relation synthesis and validation. Relations between entities were constructed using distance-based heuristics. VALUE–UNIT pairs within 12 tokens were labelled with has_unit, VALUE–PROP_TAG links were established within 260 characters, and PROP_TAG–POLYMER edges extended up to 420 characters. ADDITIVE mentions were attached to polymers if they appeared in the same sentence, while METHOD mentions within 200 characters of a VALUE triggered measured_by. After relation building, a refinement pass removed inconsistent labels. VALUEs without plausible UNITs or compatible properties were dropped. Unlinked polymers were removed unless surrounded by contextual markers (e.g., "resin", "matrix", "composite") or appearing on a strict whitelist. Each output document was validated to ensure that relation endpoints matched expected entity types.

Output format and statistics. The annotator produced JSONL files, each record containing the raw text, labelled spans (with character offsets), and relation edges. A refinement log was also written, recording the number of VALUE and UNIT labels retained after plausibility checks.

Example excerpt
Text: "The density of polystyrene was measured as 1.05 g cm^{-3} using DSC."

Labels
[4, 11, PROP_TAG] -> "density" [0]
[14, 26, POLYMER] -> "polystyrene" [1]
[43, 47, VALUE] -> "1.05" [2]
[48, 54, UNIT] -> "g cm^{-3}" [3]
[61, 64, METHOD] -> "DSC" [4]

Relations
[2, 3, "has_unit"] -> 1.05 -> g cm^{-3}
[2, 0, "describes_property"] -> 1.05 -> density
[0, 1, "about_polymer"] -> density -> polystyrene
[2, 4, "measured_by"] -> 1.05 -> DSC

A summary of refinement outcomes is provided in Table 3.5. The filtering rules removed implausible or context-free values, improving overall annotation quality.

The refinement step led to a pronounced reduction in VALUE spans, reflecting the removal of spurious numeric detections such as reference indices and unrelated numbers. UNIT spans were reduced to a lesser extent, as these tokens are typically more distinctive and less prone to false positives. The observed reduction mainly corresponds to eliminating orphan units without corresponding values, resulting in a cleaner and more reliable annotation set.

Table 3.5: Example refinement statistics aggregated across annotated corpora

Corpus	Metric	Before refinement	After refinement	Reduction
Gold	VALUE spans	4,330	2,586	40.28%
	UNIT spans	2,793	2,378	14.86%
Bulk	VALUE spans	257,748	179,224	30.47%
	UNIT spans	182,777	163,389	10.61%

3.1.5 Preprocessing

Labels. All datasets were harmonised to a unified seven-entity schema: POLYMER, ADDITIVE, PROP_TAG, VALUE, UNIT, METHOD, SOLVENT. Out-of-scope tags from the source corpora were mapped to O via a YAML mapping (e.g., POLYMER_FAMILY → POLYMER, INORGANIC/ORGANIC → ADDITIVE, PROP_NAME → PROP_TAG, PROP_VALUE/MATERIAL_AMOUNT → VALUE; all others → O).

Datasets.

- **Ramprasad abstracts**. Train/dev/test files were provided upstream, and the existing splits were retained.
- **Gold (43 docs) and Bulk (3,119 docs)**. Weak-labelled sets derived from manual pattern analysis of five seed papers. Document-level splits used a 70/10/20 (train/dev/test) ratio with a fixed random seed for reproducibility.

Tokenisation & segmentation. Preprocessing involved punctuation-based sentence splitting, followed by merging of adjacent fragments until no entity span was split across boundaries. Segments were optionally length-capped for batching but never truncated through a contiguous entity run. For transformer-based modelling, sentences were tokenised with SciBERT's WordPiece tokenizer.

Tag conversion. Source IO tags were normalised to BIO at the sentence level and then expanded to BILOU at two granularities:

- WordPiece-level BILOU for transformer fine-tuning and piece-level supervision.
- Word-level BILOU CoNLL for CRF baselines and diagnostics.

Symmetry and legality checks enforced balanced B/L counts per label and rejected illegal BILOU transitions; files that failed validation were excluded. To mitigate extreme class imbalance, segments consisting entirely of O tokens were downsampled in the training split, retaining 30% as negative context. Dev and test splits were left untouched. Because WordPiece segmentation is applied downstream of word-level segmentation, the number of segments (N) is identical across levels, so drop percentages are the same in word and piece views.

Outputs. For each corpus and split, the pipeline generated:

- *.pieces.conll (WordPiece + BILOU),
- *.words.conll (word + BILOU),
- a compact JSON/MD report summarising segment counts, drop rates, and tag inventories.

Table 3.6 summarises segment-level statistics for the three corpora. "Drop %" refers to the training sets where down-sampling of all-O sentences was applied. Segment length statistics (mean, 95th percentile, maximum) are reported across the full corpus for reference.

The overall workflow for corpus construction is summarised in Figure 3.2, covering paper harvesting, gazetteer-driven weak labelling, and preprocessing into the final train/dev/test splits.

Table 3.6: Segment-level preprocessing statistics for Ramprasad, Gold, and Bulk corpora at word and piece levels.

Corpus	Level	N	Drop % (train)	Mean	P95	Max
Ramprasad	Words	4,536	23.9	31.5	57	126
	Pieces	4,536	23.9	45.1	87	194
Gold	Words	6,576	40.9	30.8	62	1,130
	Pieces	6,576	40.9	43.4	86	2,652
Bulk	Words	442,434	48.5	31.0	64	1,111
	Pieces	442,434	48.5	44.7	92	992

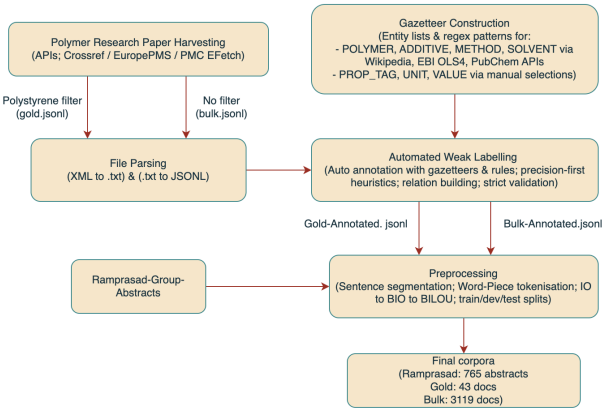

Figure 3.2: End-to-end pipeline for corpus construction

3.2 Conditional Random Fields

Conditional Random Fields (CRFs) were implemented as the first supervised baseline for sequence labelling under the unified seven-entity schema [11]. CRFs are well suited for structured prediction because they explicitly model dependencies between neighbouring labels while incorporating arbitrary token-level features. This allows them to enforce consistent BILOU sequences across entities. The aim was to establish a transparent, feature-driven benchmark before moving to recurrent and transformer-based models.

Feature engineering. Tokens were represented through a rich set of hand-crafted features. Each token was lowercased and accompanied by orthographic descriptors (word shape, compressed shape, prefixes and suffixes of length two to four). Case and digit flags were included, alongside whether the token matched curated unit regexes or contained digits. Numeric tokens were bucketed by their order of magnitude to help distinguish property scales (e.g., 10^{-3} vs 10^{3}). A context window of ± 2 tokens was added, with both surface forms and shapes. Part-of-speech tags were precomputed using spaCy (en_core_web_sm). This mix of lexical, morphological, syntactic, and domain-specific cues provided robust discriminative features for scientific text. Table 3.7 summarises the feature groups.

Training and tuning. The CRF was trained with the lbfgs algorithm (up to 250 iterations). L1 (c1) and L2 (c2) penalties were tuned via RandomizedSearchCV with 3-fold cross-validation. Weighted F1 was the optimisation metric to balance frequent and rare classes, yielding fair evaluation across diverse entity distributions.

Class imbalance. Entity distributions were uneven across corpora (e.g., METHOD absent in Ramprasad; VALUE/UNIT dominant in Bulk). Because sklearn-crfsuite lacks class weights, sentences containing rare labels were oversampled up to a cap of $3\times$. This improved recall on low-frequency classes with a modest precision trade-off.

Outputs. Predictions were evaluated on both development and held-out test sets, using strict BILOU matching. Classification reports were generated with entity-level precision, recall, and F1, providing a

Table 3.7: Feature groups and their roles in the CRF model.

Feature Group	Examples	Purpose
Lexical	Lowercased token, word prefixes/suffixes	Capture subword morphology and domain-specific cues
Orthographic	Word shape (e.g., Pppp, dd.dd), compressed shape	Normalise variants of capitalisation and mixed alphanumeric tokens
Case/Digit flags	is_title, is_upper, is_digit	Encode simple but informative lexical signals
Numeric features	Magnitude bucket (log scale)	Differentiate property values across scientific orders of magnitude
POS tags	spaCy POS categories	Provide syntactic context for entity boundaries
Domain-specific	Unit regex match, digit presence	Highlight tokens likely to be VALUEs or UNITs
Context window	± 2 token forms and shapes	Inject short-range dependencies between adjacent words

reliable baseline for subsequent neural sequence models.

While CRFs offered interpretable and efficient baselines, their reliance on manually defined features limited their ability to capture long-range dependencies and semantic variation in polymer literature. These constraints motivated the transition to recurrent and transformer-based architectures in later sections. Full CRF metrics and error breakdowns are available in the accompanying repository and are omitted here for brevity.

3.3 Bidirectional LSTM and Transformer-based Models

Neural sequence models were introduced as a progression from the feature-driven CRF baselines, with the aim of capturing richer contextual and morphological cues in polymer texts. Each successive model incorporated the strengths of its predecessors while addressing observed limitations, moving from word-only recurrent encoders to character-aware hybrids, structured sequence decoders, and ultimately transformer-based contextual embeddings. The overarching design philosophy remained precision-oriented: reliable extraction of scientific entities was prioritised over maximal recall, recognising that in technical domains, incorrect matches can be more detrimental than missing ones. The sequence of models therefore represents an iterative search for architectures that combine interpretability, precision, and robustness to the diverse vocabulary and notational patterns found in scientific articles.

3.3.1 BiLSTM baseline (Word only)

Architecture. The first neural baseline consisted of a word embedding layer, a single-layer bidirectional LSTM, and a token-wise softmax over BILOU tags. Two embedding sources were compared: 300-dimensional GloVe (trained on Common Crawl) and 200-dimensional Mat2Vec (trained on materials abstracts) [12], [13]. Tokens were mapped via a frequency-based vocabulary, with `<PAD>` at index 0 and `<UNK>` at index 1. Pretrained vectors were loaded where available; unseen tokens were randomly initialised. The encoder used a 256-dimensional BiLSTM (128 per direction). A linear projection mapped the concatenated hidden states to label logits, followed by softmax decoding. PAD suppression was enforced by masking out the `<PAD>` label at real positions.

Training and evaluation setup. Training used the Adam optimiser with a learning rate of $1e^{-3}$. Mini-batch sizes were 32 for training and 64 for validation/testing. Packed sequences ensured efficient handling of variable-length inputs. Cross-entropy loss ignored padding tokens. Early stopping with patience of three epochs was based on development weighted F1, computed with seqeval under strict BILOU evaluation. Multiple random seeds were used with deterministic cuDNN settings to control variance. A diagnostic check confirmed that the model never emitted `<PAD>` labels on real tokens after

training.

Motivation. This baseline served to establish a transparent neural benchmark against the feature-engineered CRFs and to test whether contextual sequence modelling at the word level improved performance on scientific text. Comparing GloVe with Mat2Vec embeddings also provided an initial assessment of whether broad lexical coverage or domain-specific pretraining was more advantageous.

Observed limitations. Performance on frequent labels such as PROP_TAG and UNIT was strong, but the model was brittle to morphological variants and unseen abbreviations. Polymers expressed in shorthand (e.g., PS, PMMA) were often missed outside familiar contexts, and units with non-standard formatting were inconsistently tagged. Although Mat2Vec offered a domain origin, its vocabulary coverage was smaller than GloVe's, limiting its effectiveness. GloVe embeddings provided more stable results, suggesting that lexical coverage outweighed domain specialisation at this stage.

Transition. The weaknesses of the word-only encoder emphasised the need for subword modelling to capture rare forms and morphological cues. This motivated the introduction of character-level encoders in the next set of models (Sections 3.3.2 & 3.3.3), while retaining GloVe embeddings as the stronger word-level backbone.

3.3.2 Character-aware BiLSTMs (CharCNN vs CharBiLSTM)

Architecture. To address the vocabulary and morphology limitations of word-only BiLSTMs, two character-aware extensions were implemented. In both cases, each token was decomposed into a sequence of characters drawn from a fixed vocabulary of 467 symbols, including padding and unknown markers. Characters were embedded into a 30-dimensional vector space, and word-level embeddings from 300-dimensional GloVe vectors were concatenated with the character representations before being passed into a word-level BiLSTM encoder of hidden size 256.

The first variant (CharCNN + BiLSTM) employed a convolutional encoder at the character level. A one-dimensional convolution with kernel sizes of three to five was applied over the character embeddings, followed by max pooling across the sequence length to obtain fixed-size vectors. This design emphasised local n-gram patterns such as "poly-", "-ate", or "styrene".

The second variant (CharBiLSTM + BiLSTM) replaced the convolutional encoder with a recurrent one. Character sequences were processed using a bidirectional LSTM with hidden sizes of 25 or 50 per direction, and the final forward and backward states were concatenated to form subword embeddings. This encoder preserved the sequential order of characters, allowing modelling of long acronyms and multi-symbol units such as "g·mol^{-1}" or "W·m^{-1}·K^{-1}".

Training setup. Both variants were trained with the Adam optimiser at a learning rate of 1e^{-3}, using mini-batches of 32 for training and 64 for evaluation. Gradient clipping at 5.0 was applied to stabilise training. Cross-entropy loss was used, masking padding positions. Early stopping with a patience of three epochs monitored development set weighted F1. For the CharCNN encoder, a grid search was conducted over filter sizes {3, 4, 5} and output dimensions {25, 50}, while for the CharBiLSTM encoder hidden sizes of {25, 50} were screened. Each configuration was repeated across multiple random seeds to mitigate variance, and the best-performing setup was selected for test evaluation.

Motivation. The motivation for introducing character-level encoders was to improve generalisation beyond fixed word embeddings. Scientific texts contain numerous abbreviations, morphological variants, and irregular unit notations. While word embeddings capture semantic similarity, they are limited when encountering unseen tokens. The character encoders aimed to provide robustness by encoding subword structure, complementing the distributional information from GloVe.

Comparative observations. The CharCNN encoder provided efficient and robust handling of common prefixes, suffixes, and short n-gram cues. However, its max-pooling step discarded sequential information, making it less suited for long acronyms and systematic chemical names where order is critical. By contrast, the CharBiLSTM preserved character order and therefore improved representation of long, complex tokens. This came at the cost of substantially higher training time and parameter overhead, and performance stability was weaker on smaller corpora.

Transition. Despite its expressive advantages, the CharBiLSTM was not chosen for the progressive model design. The CharCNN variant provided a more favourable balance of robustness, efficiency, and training stability, particularly under the constraints of large weakly labelled corpora. In line with the

precision-first design philosophy, CharCNN avoided the overfitting risks and computational overhead associated with CharBiLSTM, while still capturing the key morphological cues required for scientific text. As a result, the CharCNN + BiLSTM design was retained as the character-aware foundation for the BiLSTM+CRF architecture (Section 3.3.3).

3.3.3 BiLSTM + CRF (CharCNN)

Architecture. The third progressive variant extended the CharCNN + BiLSTM encoder with a Conditional Random Field (CRF) decoding layer [14], [15]. Tokens were represented through the concatenation of 300-dimensional GloVe vectors and character-level embeddings produced by the convolutional encoder described above. The combined word–character embeddings were passed into a bidirectional LSTM with hidden size 256, yielding contextualised token representations. A linear projection mapped these to emission scores, which were then fed into a CRF layer. Unlike the independent softmax decoder of earlier models, the CRF estimated the probability of the full label sequence and enforced legal BILOU transitions. Invalid outputs, such as an "I-UNIT" without a preceding "B-UNIT", were forbidden by transition constraints. Padding was also suppressed at real positions to prevent degenerate predictions.

Training setup. The BiLSTM+CRF was trained using the Adam optimiser with a learning rate of 5×10^{-4}. The training objective was the negative log-likelihood of the gold label sequence under the CRF. Gradient clipping at 5.0 was applied, and early stopping with patience of three epochs was triggered if no improvement was observed on the development set weighted F1. Transition matrices were initialised with constraints reflecting the BILOU tagging scheme: start states disallowed inside or last tags, end states disallowed beginning or inside tags, and transitions to and from padding were forbidden.

Training was repeated with multiple random seeds under deterministic cuDNN settings to reduce variance, and the best run was selected on the basis of weighted F1 on the development set.

Motivation. The motivation for introducing the CRF decoder was to resolve a weakness observed in earlier BiLSTM models. Although token-level predictions were accurate, entity spans were sometimes fragmented, particularly for long property names such as "glass transition temperature" or systematic polymer names. In scientific information extraction, span integrity is crucial, since broken entity boundaries can invalidate downstream property–value mappings. The CRF provided a structured decoding mechanism that enforced global consistency, ensuring that predicted sequences formed complete entities rather than partial fragments.

Observed limitations. The CRF improved boundary consistency and reduced span fragmentation across properties, polymers, and multi-token units. Nevertheless, it did not address the data sparsity affecting low-frequency classes such as VALUE and UNIT, where recall remained constrained by limited positive examples in training. The CRF also introduced additional computational cost during training and decoding, although this overhead was still moderate compared with transformer-based models.

Transition. This architecture combined character-aware embeddings with globally consistent decoding, producing a stable and high-precision model. However, it still relied on static GloVe embeddings, which lacked the capacity to capture domain-wide contextual variation and long-range dependencies. To explore whether broader contextual representations could offer further improvements, the next stage incorporated SciBERT embeddings, a widely used scientific transformer, as a reference point for transformer-level modelling in this domain.

3.3.4 SciBERT-based models

Architecture. The final set of experiments introduced transformer-based contextual embeddings through SciBERT [17]. Two configurations were considered. In the frozen setup, contextual embeddings were extracted from SciBERT and pooled at the word level before being passed into a BiLSTM-CRF decoder. In the fine-tuned setup, the top four encoder layers of SciBERT were unfrozen and optimised jointly with the BiLSTM-CRF sequence tagger. This provided a direct comparison between lightweight feature extraction and full contextual adaptation within the pipeline already established with GloVe embeddings and character encoders.

Training setup. Frozen feature experiments employed a learning rate of 1e-3 for the BiLSTM and CRF layers, while SciBERT parameters remained fixed. Fine-tuning used AdamW with learning rates in the

range of 2e-5 to 3e-4, supported by layer-wise decay and linear warm-up scheduling. Gradient clipping at 5.0 and early stopping after three epochs without improvement were applied. Batch sizes were reduced to 8–16 to manage GPU memory constraints. Models were trained under multiple seeds with deterministic settings, and the best checkpoint was selected based on weighted F1 on the development set.

Motivation. SciBERT was introduced not as the expected best-performing model but as a scientifically recognised baseline for contextualised embeddings. Its vocabulary and training corpus cover a wide range of scientific literature, making it a reproducible reference point for assessing whether transformer-level representations add measurable value over lighter recurrent encoders in polymer property extraction. Domain-specialised alternatives such as MatSciBERT or PolyBERT were acknowledged in the literature review, but their heavier memory requirements and less mature tooling made them impractical within the project's computational and time constraints. SciBERT thus offered a controlled, well-supported benchmark that enabled a rigorous comparison without conflating multiple changes in model scope and pretraining.

Observed limitations. SciBERT improved over word-only BiLSTMs on both the Gold and Ramprasad corpora, confirming that contextual embeddings added meaningful signals even in weakly labelled settings. However, the gains did not surpass the stronger character-aware BiLSTM+CRF variant, which remained more stable and resource-efficient. Fine-tuning SciBERT increased accuracy but introduced variance across runs, while frozen feature extraction was consistent but plateaued below the recurrent baselines. Computational overhead was considerable: training times were significantly longer, and GPU memory demand prevented large-scale runs on the Bulk corpus. Given that the BiLSTM+CRF already achieved near-ceiling performance on Bulk at a fraction of the cost, extending SciBERT to that corpus was deprioritised.

Transition. SciBERT validated the role of transformer-based contextual embeddings by demonstrating competitive results on small scientific corpora, but it also highlighted that such models are not automatically superior under weak supervision and constrained resources. Character-aware BiLSTM+CRF remained the more practical choice in this project, balancing robustness and efficiency. Nonetheless, the results suggested that domain-specialised transformers such as MatSciBERT or PolyBERT could offer further advantages, provided that larger curated corpora and sufficient resources are available for stable fine-tuning.

Together, these variants established a progression from transparent, feature-based models to increasingly expressive neural architectures. The evaluation results presented in the next section highlight how these design choices translated into empirical performance across the different corpora.

.

Chapter 4
Results and Discussion

4.1 Overall Performance

4.1.1 Evaluation Framework

All models were evaluated under the strict BILOU tagging scheme using the `seqeval` scorer [14], [15]. Crucially, the classification reports excluded the "O" (outside) label from all calculations, ensuring that only entity spans influenced the metrics. Three complementary metrics were reported:

- **Weighted F1** — sensitive to label frequency, reflecting aggregate accuracy across all classes.
- **Macro F1** — each class contributed equally, highlighting weaknesses on rare or imbalanced entities.
- **Per-label F1** — entity-wise scores, exposing which categories benefited from specific model features.

This combination provided a balanced view: weighted F1 established the overall benchmark, macro F1 revealed the effect of class imbalance, and per-label scores showed where design choices most influenced outcomes.

4.1.2 Corpus-level comparison

Table 4.1 summarises the best-performing configuration per corpus, selected on the basis of weighted F1 on the development set.

Table 4.1: Best model performance per corpus (test set).

Corpus	Best Model	Weighted F1	Macro F1	Micro F1
Bulk (3,119 docs, varied polymers)	BiLSTM+CRF (CharCNN)	0.974	0.969	0.974
Gold (43 docs, polystyrene-specific)	BiLSTM+CRF (CharCNN)	0.866	0.709	0.883
Ramprasad (≈765 abstracts, mixed polymers)	SciBERT (fine-tuned, top-4)	0.686	0.639	0.690

Performance followed clear corpus-level patterns. On the Bulk corpus, character-aware BiLSTM+CRF reached near-ceiling scores, with weighted and macro F1 both above 0.96. On the Gold corpus, despite being weakly labelled and small in scale, the model still achieved ∼0.87 weighted F1, demonstrating that weak supervision can yield reliable outcomes when applied to a focused single-polymer domain. By contrast, the Ramprasad corpus (small and heterogeneous across many polymers) remained more challenging, where fine-tuned SciBERT slightly outperformed recurrent models but absolute scores stayed modest.

Transfer Learning. A BiLSTM+CRF (CharCNN) model trained on the Bulk corpus was evaluated on the Gold train/dev/test splits without any fine tuning on Gold under the same strict BILOU protocol (O excluded), with hyperparameters fixed from Bulk. It achieved weighted/macro F1 of 0.96/0.96 (train), 0.98/0.96 (dev), and 0.97/0.96 (test). Despite only two Gold documents being present in Bulk and forty-one being unseen, performance remained consistently high, indicating effective transfer from large weakly labelled data to the smaller target corpus; detailed reports are available in the repository.

4.1.3 Model family trends

Recurrent baselines (BiLSTMs). Across all corpora, BiLSTM architectures produced competitive and reliable results. Word-only baselines struggled on morphology-sensitive entities such as polymers and values, while the addition of character encoders improved robustness to unseen forms. The BiLSTM+CRF variant provided the strongest balance of precision and boundary consistency, particularly in the Gold corpus, where its stability highlighted that weak supervision was sufficient to train effective models when domain variation was controlled.

Transformer models (SciBERT). SciBERT showed mixed performance. On the Gold corpus, fine-tuning occasionally surpassed BiLSTMs and delivered higher macro F1, but the gains were not large enough to justify the substantial computational overhead. On the Ramprasad corpus, its contextual embeddings yielded a clearer advantage over recurrent baselines, though absolute scores remained lower than those of Gold due to the heterogeneous mix of polymers. On Bulk, SciBERT was not attempted because BiLSTM models had already saturated performance at a fraction of the computational cost.

Cross-corpus insight. Taken together, the results indicated that data composition was as important as size. The Bulk corpus confirmed that large, weakly labelled collections spanning diverse polymers allowed BiLSTM+CRF to reach near-perfect precision. The Gold corpus showed that even a small, polystyrene-specific dataset could support robust modelling under weak supervision, provided the domain was coherent. By contrast, the Ramprasad abstracts, though manually labelled, were too small and varied to establish consistent trends, limiting the gains from contextual embeddings.

4.1.4 Per-label analysis

Figure 4.1 presents a combined heatmap of per-entity F1 scores for the best-performing model on each corpus. This overview highlights consistent strengths across common entities and exposes the categories that remained most challenging.

On the Bulk corpus, all entity types scored above 0.90 F1, with PROP_TAG and SOLVENT exceeding 0.99. The only comparatively weaker class was POLYMER (∼0.88 F1), which reflected the high morphological diversity of polymer names spanning multiple families. On the Gold corpus, METHOD and PROP_TAG remained strong (above 0.92 F1), while VALUE and UNIT declined into the 0.77–0.83 range. POLYMER fell below 0.40, a consequence of sparse and uneven annotation of polymer mentions in the polystyrene-specific set.

On the Ramprasad corpus, fine-tuned SciBERT improved the recognition of longer entities, particularly VALUE and UNIT, relative to recurrent baselines. However, POLYMER and ADDITIVE remained difficult, consistent with their limited representation in the manually labelled abstracts.

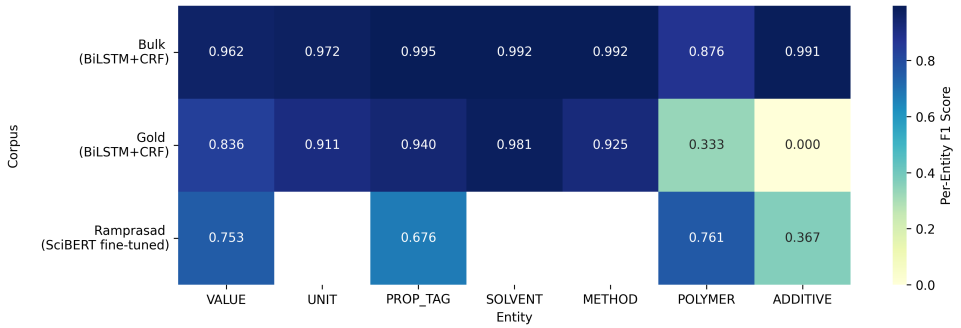

Figure 4.1: Heatmap of per-entity F1 scores for best-performing models on Bulk (BiLSTM+CRF), Gold (BiLSTM+CRF), and Ramprasad (SciBERT fine-tuned)

4.1.5 Interpretation and transition

The comparative evaluation produced three insights. First, large-scale weak supervision proved highly effective when applied to diverse corpora, as demonstrated by the near-ceiling Bulk results. Second,

when data were limited, domain focus mattered: the Gold corpus outperformed the more heterogeneous Ramprasad dataset despite being weakly labelled, confirming that weak supervision can be both efficient and reliable when applied within a constrained scope. Third, while transformers such as SciBERT added contextual depth and improved macro F1 in some cases, their additional cost did not translate into consistent value beyond BiLSTM+CRF under the current data conditions.

These findings established BiLSTM+CRF as the most practical backbone for subsequent experiments. The next sections investigate per-entity behaviours and error categories in detail, illustrating where models succeeded, where they failed, and what these patterns reveal about polymer information extraction.

4.2 Model-wise Comparison

4.2.1 Rationale for model-level evaluation

Whereas Section 4.1 established corpus-level outcomes, this section compares all model variants directly to trace how successive design choices influenced performance. The evaluation considered both weighted and macro F1, reported side by side for each model. Weighted F1 reflected aggregate stability across frequent classes, while macro F1 ensured that weaknesses on rare entities were not masked by class imbalance. Taken together, these measures provided a balanced assessment of whether architectural refinements genuinely improved robustness.

4.2.2 Comparative performance across models

Figures 4.2–4.4 present bar charts of weighted and macro F1 for every model tested on the Bulk, Gold, and Ramprasad corpora. Each architecture is displayed with two bars, one representing weighted F1 and the other macro F1. This visualisation allowed direct comparison of how word-only baselines, character-aware extensions, CRF decoding, and transformer embeddings contributed to overall reliability.

On the Bulk corpus, word-only BiLSTMs achieved strong weighted F1 but revealed gaps in macro F1, especially on polymers and values. Introducing character encoders reduced this disparity: CharCNN and CharBiLSTM lifted macro F1 closer to weighted levels, reflecting improved handling of rare or morphologically complex tokens. Adding a CRF on top of the character-aware BiLSTM produced the highest scores, with weighted and macro F1 both exceeding 0.96. This pattern confirmed that each architectural refinement contributed incrementally, and by the time of the BiLSTM+CRF, further gains from more complex models were marginal.

On the Gold corpus, differences between models were more pronounced. Word-only BiLSTMs achieved weighted F1 around 0.83–0.85 but lagged in macro F1, indicating fragility on rare entities. CharCNN and CharBiLSTM offered substantial improvements, especially in macro F1, showing that subword encoders were crucial when data were limited. The BiLSTM+CRF reached the most stable balance, with weighted F1 of 0.87 and macro F1 above 0.70. Fine-tuned SciBERT occasionally surpassed BiLSTM+CRF in macro F1, reflecting its capacity to capture rare categories, although its variance across seeds limited consistency. Frozen SciBERT, in contrast, performed below recurrent models, demonstrating that without adaptation contextual embeddings did not add value in this setting.

On the Ramprasad corpus, the situation reversed. Here SciBERT fine-tuned clearly outperformed recurrent baselines, with weighted and macro F1 in the high 0.68 range compared with 0.62–0.65 for BiLSTMs. Character-aware variants offered some gains over word-only baselines, but the effect was modest. The transformer's contextual embeddings were more effective on this small but carefully labelled corpus, particularly for long multi-token values and units. However, absolute scores remained lower than in Gold, reflecting the heterogeneous polymer coverage of the Ramprasad abstracts.

4.2.3 Interpretation

The cross-model comparison yielded three consistent insights. First, character-aware encoders provided reliable improvements over word-only BiLSTMs, especially in smaller or weakly supervised datasets where morphological cues were critical. Second, CRF decoding delivered further gains by enforcing span integrity, establishing the BiLSTM+CRF as the strongest recurrent architecture across both Bulk

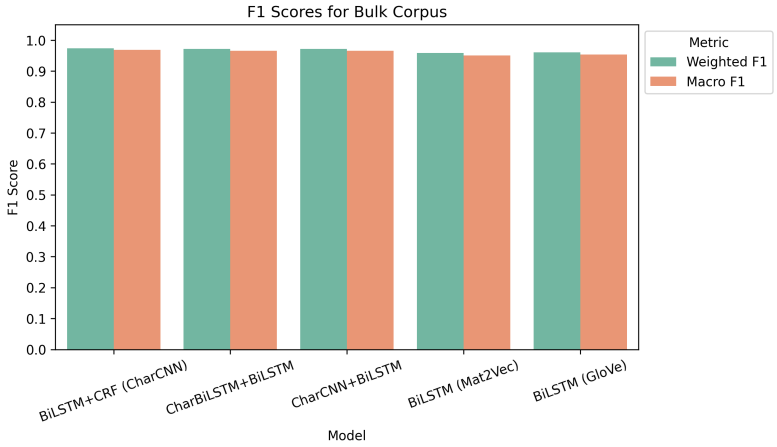

Figure 4.2: Weighted and macro F1 by model on Bulk corpus

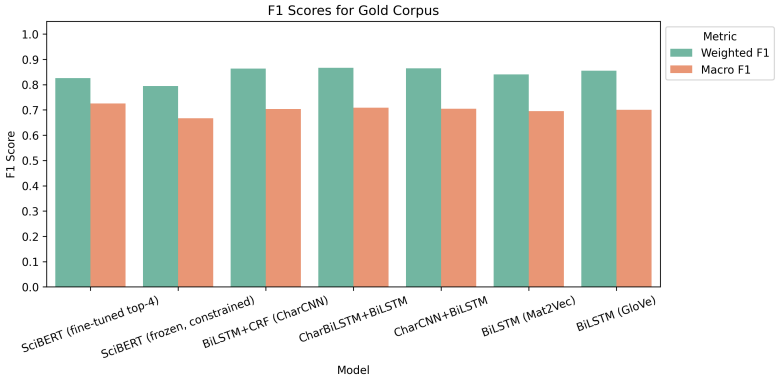

Figure 4.3: Weighted and macro F1 by model on Gold corpus

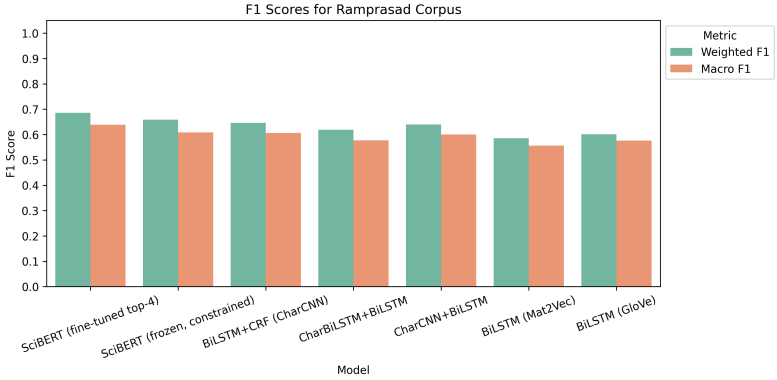

Figure 4.4: Weighted and macro F1 by model on Ramprasad corpus

and Gold. Third, transformer embeddings showed clear potential when applied to small but clean datasets, as in Ramprasad, but offered less value under weak supervision and required considerably more computational resources.

Overall, the evidence positioned BiLSTM+CRF as the most effective and efficient backbone for large-scale polymer information extraction, while contextual transformers such as SciBERT remained promising for future work where domain-specialised pretraining and stronger supervision are available. The next section examines per-entity error categories to clarify where models succeeded and where limitations persisted.

4.3 Error Analysis

While overall performance metrics established the relative strengths of different architectures, they did not reveal the precise nature of the remaining errors. To address this, errors were categorised both at the level of individual entities and at the level of sentence outcomes. The aim was to distinguish whether failures arose from boundary inconsistencies, incorrect type assignments, or complete omissions, and to quantify how often predictions were partially or entirely correct within their original sentence contexts.

4.3.1 Distribution of error types across entities

Figures 4.5–4.7 illustrate the distribution of boundary and type errors across entities for each corpus. These analyses highlighted consistent patterns.

On the Bulk corpus, where the BiLSTM+CRF achieved near-ceiling scores, residual errors were almost entirely boundary-related. METHOD and SOLVENT recorded the highest counts, with errors concentrated in multi-token spans such as analytical techniques and solvent names. POLYMER errors were fewer in number but still showed truncation of systematic names. VALUE, UNIT, and ADDITIVE were generally stable, with only scattered boundary slips, and type errors remained rare across all categories.

In the Gold corpus, the error distribution was minimal, reflecting both the smaller dataset and its polystyrene-specific focus. METHOD accounted for most issues with 17 boundary errors, typically involving longer multi-word phrases. PROP_TAG contributed 3 boundary and 1 type error, usually in descriptive property expressions. All other entities, including POLYMER, VALUE, UNIT, ADDITIVE, and SOLVENT, were extracted without type error. This outcome suggested that, within a narrow domain, boundary segmentation rather than type confusion remained the dominant weakness.

In the Ramprasad corpus, SciBERT fine-tuned exhibited a larger proportion of type errors. POLYMER was particularly affected, with both truncations and misclassifications, while ADDITIVE showed frequent type confusions against other chemical categories. VALUE performed moderately well but often lost boundary precision in longer numeric–unit expressions. PROP_TAG also carried a mix of type and boundary errors, reflecting the heterogeneity of property descriptors across the abstracts. METHOD and UNIT, by contrast, were consistently reliable with zero recorded errors.

4.3.2 Sentence-level error buckets

In addition to entity-level analysis, sentences were grouped into four buckets: perfect predictions, mixed predictions where some entities were correct and others incorrect, completely missed predictions, and spurious outputs where entities were predicted despite not being present in the gold annotations. Figures 4.8 summarise the proportions of these buckets for each corpus.

On the Bulk corpus, more than ninety-seven percent of sentences fell into the perfect category, reflecting the stability of BiLSTM+CRF when trained on large and diverse data. Mixed outcomes accounted for less than three percent, and spurious predictions were negligible. The few errors that did occur almost exclusively involved boundary truncations of complex polymer names.

In the Gold corpus, approximately eighty to eighty-five percent of sentences were perfectly predicted, with the remainder dominated by mixed cases. These often-involved correct identification of a property tag but omission of the associated value or unit, or incomplete capture of a polymer span. Missed and

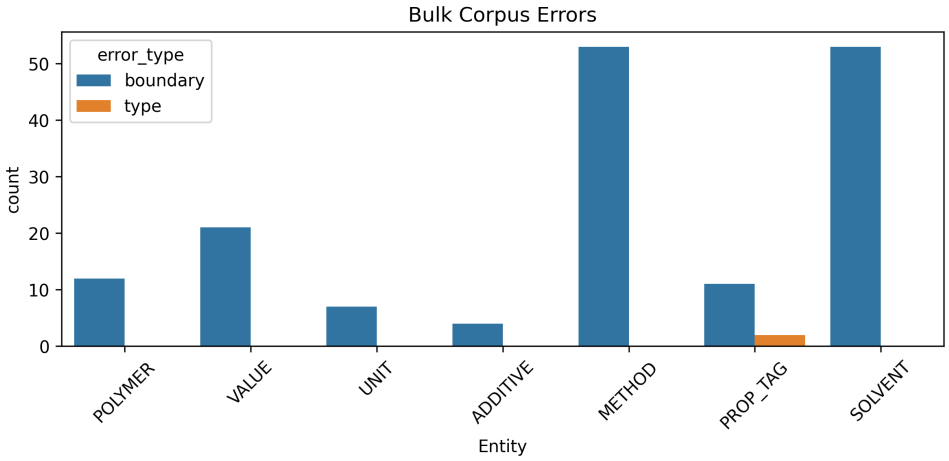

Figure 4.5: Distribution of error types by entity on Bulk corpus

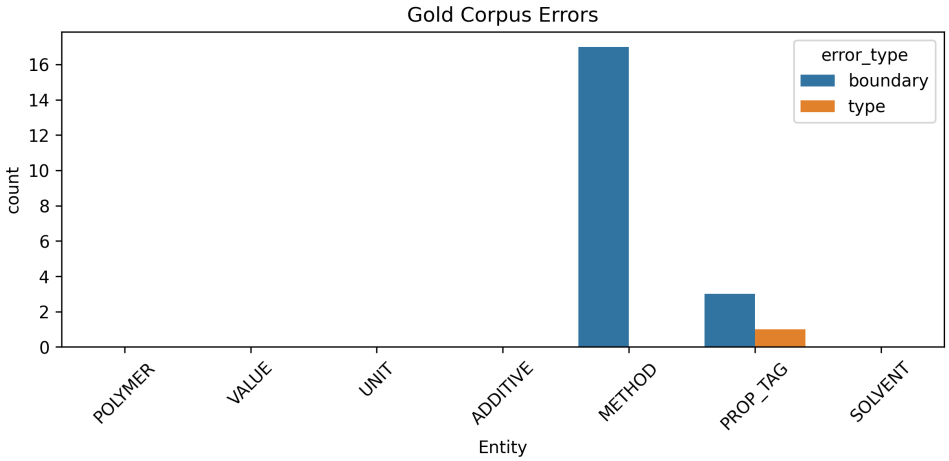

Figure 4.6: Distribution of error types by entity on Gold corpus

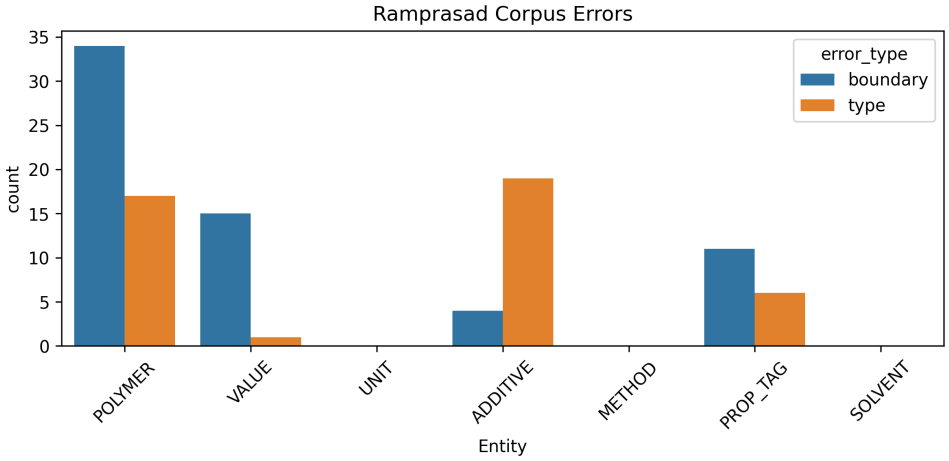

Figure 4.7: Distribution of error types by entity on Ramprasad corpus

spurious cases were rare, indicating that the main challenge was partial coverage rather than outright failure.

The Ramprasad corpus presented a different profile. Only about half of sentences were predicted perfectly, with a large share classified as mixed. These typically involved partial capture of property–value pairs or inconsistent labelling of additives and polymers. Missed sentences and spurious predictions together accounted for ten to fifteen percent of the data, highlighting the instability of model behaviour in this setting.

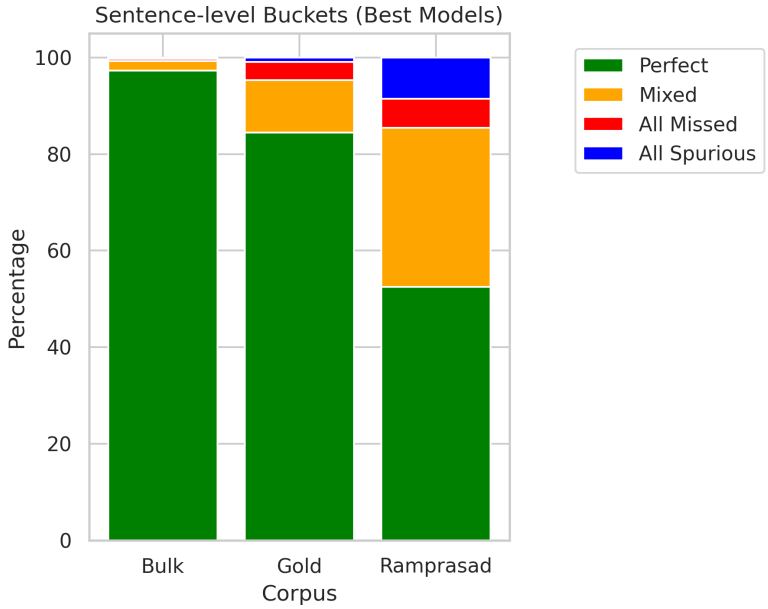

Figure 4.8: Sentence-level prediction buckets for best-performing models on Bulk (BiLSTM+CRF), Gold (BiLSTM+CRF), and Ramprasad (SciBERT fine-tuned)

4.3.3 Interpretation

The error analysis clarified three points. First, boundary errors were the dominant failure mode across all corpora, particularly for morphologically complex or abbreviated polymer names. Second, type errors were concentrated in VALUE and UNIT classes under conditions of annotation sparsity, confirming that numeric forms and unusual unit expressions remained challenging for both recurrent and transformer models. Third, sentence-level analysis revealed that large and domain-focused corpora not only improved overall scores but also minimised mixed and spurious outcomes, whereas smaller heterogeneous data sources amplified instability.

Together these findings reinforced the precision-first philosophy of this work: although overall scores were high, reliable boundary integrity and reduced mixed cases were equally important for downstream applications. The next section builds on these aggregate findings by examining concrete sentence examples that illustrate the strengths and limitations of the models in practice.

4.4 Qualitative Examples

Quantitative evaluation established performance patterns, but sentence-level inspection provided richer insight into how models handled the complexities of polymer texts. Examples were grouped into four categories: (i) perfect predictions, where entities were captured exactly as annotated; (ii) boundary errors, where spans were truncated or extended; (iii) omissions, where one or more entities in a relation were missed; and (iv) spurious predictions, where the model introduced entities not supported by the text. This analysis was applied across Bulk, Gold, and Ramprasad corpora.

4.4.1 Bulk corpus

Perfect predictions

1. "The glass transition temperature of polystyrene was 103 °C." — The property phrase (glass transition temperature), the numeric value (103), and the unit (°C) were all correctly identified.This illustrates the model's ability to capture complete property–value–unit triplets, which are central to polymer information extraction.

2. "Differential scanning calorimetry was used to determine thermal stability." — The method term was tagged precisely as METHODThe absence of surrounding numerical or polymer references ensured that the entity was unambiguous, and the model exploited this clarity effectively.

Boundary errors

1. "Poly(methyl methacrylate) blended with polystyrene was used as the matrix." — The correct label was *Poly(methyl methacrylate)*, yet the model often truncated the span to *methyl methacrylate*. This error highlights a weakness in recognising full polymer names that include functional prefixes such as *Poly(*.

2. "Poly(styrene-co-acrylonitrile) composites were tested for thermal stability." — Instead of capturing the full copolymer, the model occasionally split the entity around the hyphen, producing partial spans. Such errors reflect the structural complexity of systematic polymer nomenclature.

Omissions

1. "Elastic modulus increased to 1.5 GPa after crosslinking." — While the property term *elastic modulus* was tagged correctly, the corresponding value and unit were not captured.

2. "The density of the nanocomposite was 1.2 g cm^{-3}." — In some instances, the model dropped the numeric value *1.2* while retaining *g* cm^{-3}.

Spurious predictions

1. "Samples were cast into thin films." — A false UNIT tag was introduced despite the absence of any measurement.

2. "The polymerisation was initiated under nitrogen atmosphere." — The token *nitrogen* was misclassified as a solvent.

4.4.2 Gold corpus

Perfect predictions

1. "The refractive index of polystyrene was found to be 1.59." — Both the property phrase *refractive index* and the numeric value *1.59* were correctly tagged.

2. "PS samples showed a glass transition temperature of 103 °C." — The abbreviation *PS* was correctly identified as a polymer, and the property–value–unit triplet was fully captured.

Boundary errors

1. "The polystyrene homopolymer exhibited higher Tg than blends." — The model often reduced the phrase to *polystyrene*, losing the modifier *homopolymer*.

2. "Molecular weight distribution of polystyrene fractions was measured." — The correct span *molecular weight distribution* was fragmented into *molecular weight* and *distribution*.

Omissions

1. "PS films had a density of 1.05 g cm^{-3}." — The property label *density* was correctly tagged, but the value and unit were sometimes omitted.

2. "The samples showed reduced thermal stability." — The property phrase *thermal stability* was inconsistently tagged.

Spurious prediction

1. "The polymer films were transparent under visible light." — A false VALUE tag was introduced for *transparent*.

2. "Samples were dried in air before testing." — *Air* was incorrectly tagged as a solvent.

4.4.3 Ramprasad corpus

Perfect predictions

1. "Thermal conductivity was reported in W m^{-1} K^{-1}." — Both the property and the multi-symbol unit were correctly tagged.

2. "Molecular weights of the copolymers were determined by GPC." — The method *GPC* was accurately extracted.

Boundary errors

1. "Blends of poly(ethylene oxide) with lithium salts were characterised by DSC." — The polymer was identified correctly, but the phrase *lithium salts* was truncated.

2. "Poly(vinyl alcohol) solutions were studied at different concentrations." — The span was sometimes reduced to *vinyl alcohol*, dropping the prefix *poly*.

Omissions

1. "The samples exhibited a refractive index of 1.52." — The property was captured, but the value *1.52* was omitted.

2. "Molecular weights were reported for the copolymer." — The property tag was detected, but numeric values were missing.

Spurious predictions

1. "Molecular weight distribution was determined." — A false VALUE tag was generated without a number.

2. "The samples were annealed prior to testing." — The process term *annealed* was misclassified as a property.

4.4.4 Cross-model progression examples

To illustrate how architectural changes influenced performance, the same sentences were compared across multiple models. These cases highlight why BiLSTM+CRF was ultimately the most effective backbone.

Bulk corpus (varied polymers)
Sentence: "Poly(methyl methacrylate) blended with polystyrene was used as the matrix."

- **Word-only BiLSTM:** Truncated *Poly(methyl methacrylate)* to *methyl methacrylate* and omitted *polystyrene*, resulting in incomplete polymer identification.
- **CharCNN+BiLSTM:** Detected both polymers, but boundaries around *Poly(methyl methacrylate)* were sometimes cut short.

- **BiLSTM+CRF:** Correctly captured both polymers with full spans, resolving truncation errors and ensuring reliable entity recognition.

Gold corpus (polystyrene-specific)
Sentence: "PS samples showed a glass transition temperature of 103 °C."

- **Word-only BiLSTM:** Correctly tagged *glass transition temperature* but missed *PS* as a polymer, leading to an incomplete property–value–unit triplet.
- **CharCNN+BiLSTM:** Detected *PS* as a polymer and the numeric value *103*, but occasionally dropped the unit °C, leaving a partial triplet.
- **BiLSTM+CRF:** Captured the full set of entities (*PS*, *glass transition temperature*, *103 °C*) with span integrity, producing a complete and consistent triplet.

Ramprasad corpus (mixed polymer abstracts)
Sentence: "Thermal conductivity was reported in W m^{-1} K^{-1}."

- **Word-only BiLSTM:** Detected *thermal conductivity* but missed the complex unit, weakening the property–value link.
- **CharCNN+BiLSTM:** Captured the unit partially, often truncating after $W\ m^{-1}$.
- **SciBERT (fine-tuned):** Correctly extracted both *thermal conductivity* and the entire unit $W\ m^{-1}\ K^{-1}$, showing that contextual embeddings were more effective for long, composite symbols.

4.4.5 Interpretation

The balance of correct and erroneous examples clarified model behaviour. Perfect predictions confirmed that BiLSTM+CRF handled property–value–unit triplets with high reliability, particularly in Bulk where training scale was maximised. Boundary errors exposed the difficulty of complex polymer names and acronyms, which remain challenging even for contextual models. Omissions highlighted weaknesses in linking values and units to properties, a recurring bottleneck in smaller corpora. Spurious predictions illustrated overgeneralisation, where frequent co-occurrence patterns led to false positives.

Together, these cases confirmed that BiLSTM+CRF provided the most stable backbone for polymer information extraction. It captured standardised expressions with precision while avoiding the excessive variance and computational cost of transformer-based approaches. Contextual models such as SciBERT improved specific cases, particularly for complex units and long property phrases, but their advantages were limited in weakly labelled settings.

4.5 Results Summary

The results across Bulk, Gold, and Ramprasad corpora collectively addressed the research aims and clarified the factors that most shaped extraction quality. A clear progression was observed in model performance. Word-only BiLSTMs established a basic neural benchmark but struggled with morphology and abbreviations. Adding character-level encoders improved robustness to unseen variants, and the BiLSTM+CRF consistently delivered the most stable and precise predictions by enforcing span consistency. Transformer-based contextual embeddings provided selective gains in certain settings but introduced instability and higher computational cost, limiting their practical utility in this project.

Corpus composition proved to be as decisive as architecture. The Bulk corpus demonstrated that large, weakly labelled collections enabled near-ceiling accuracy with recurrent models. The Gold corpus, although much smaller, showed that weak supervision can be effective when applied to a focused single-polymer domain, producing outcomes superior to the heterogeneous Ramprasad abstracts. These findings confirmed that annotation strategy and domain specificity were as important as model design for reliable performance.

Additional experiments with class weighting and sentence oversampling were also explored to address imbalance, but neither produced consistent improvements. Weighted training matched the unweighted baseline, while oversampling reduced stability. These results suggested that architectural refinements and robust contextual features contributed more to reliability than data rebalancing heuristics.

In summary, the BiLSTM+CRF architecture emerged as the most practical backbone for polymer information extraction, balancing precision, stability, and efficiency across diverse corpora. SciBERT demonstrated the potential of contextual embeddings but did not offer sufficient advantage in weakly supervised conditions to justify its cost. These outcomes establish a foundation for the discussion in Chapter 5, where their broader implications for scientific text mining, precision-first extraction, and future domain-specific transformer models are considered.

Chapter 5

Conclusion and Future Work

5.1 Revisiting Research Questions and Aims

This project set out to design and evaluate a precision-first pipeline for extracting polymer property information from scientific literature. The overarching aim was to test whether weakly labelled corpora could support reliable named entity recognition when paired with appropriate model architectures. The five research questions introduced in Chapter 1 are revisited here.

RQ1: How can weakly labelled full-text corpora be constructed in a way that balances coverage and precision?

The study showed that weak supervision is viable if corpus composition is carefully controlled. The Bulk corpus, built from thousands of weakly labelled papers spanning diverse polymers, delivered near-ceiling performance when scale compensated for noise. The Gold corpus, although much smaller, achieved competitive scores because it was restricted to a single polymer family (polystyrene). By contrast, the Ramprasad corpus of mixed abstracts remained challenging, indicating that small but heterogeneous datasets dilute the benefits of weak supervision. The findings suggest that either scale or domain focus is required for weak supervision to be effective.

RQ2: Does expanding the schema with additional entity types and relations improve the contextual richness of extracted information?

Expanding the schema beyond the core polymer–property–value triad to include METHOD, ADDITIVE, and SOLVENT increased the descriptive coverage of extracted information. This allowed outputs to capture experimental setups and compositional factors alongside core property data. The trade-off was that lower-frequency entities such as ADDITIVE and SOLVENT were more difficult to learn reliably, particularly in the smaller corpora. Nevertheless, their inclusion proved valuable for contextual completeness, demonstrating that schema enrichment contributes to richer representations even if some labels remain harder to model.

RQ3: What performance gains can be achieved by moving from feature-based CRF models to recurrent and transformer-based architectures?

The results confirmed that neural architectures consistently outperformed feature-engineered CRFs. Word-only BiLSTMs provided a stronger baseline than CRFs, while character-aware BiLSTM variants further improved robustness to morphology and unseen tokens. The BiLSTM+CRF achieved the most stable and precise outcomes overall, particularly on the Bulk and Gold corpora. Transformer-based SciBERT added contextual depth and improved handling of complex notations in some cases, but the gains were inconsistent under weak supervision and came at significant computational cost.

RQ4: How do models trained on abstracts compare to those trained on full papers in terms of generalisation and strict F1 performance?

Models trained on the full-text Bulk and Gold corpora achieved stronger generalisation than those trained on the abstract-only Ramprasad set. The abstract corpus lacked consistent contextual cues, leading to poorer recall and noisier predictions despite using SciBERT. By contrast, full papers provided richer co-occurrence patterns between polymers, properties, and values, enabling the BiLSTM+CRF to generalise effectively even under weak labelling. This confirmed that full-text corpora are more conducive to stable extraction than abstract-only datasets. Notably, a model trained on the Bulk corpus transferred to Gold without any fine tuning and with hyperparameters fixed from Bulk, maintaining high weighted and macro F1 across the train/dev/test splits, which evidences robust cross-corpus generalisation from large weakly labelled data.

RQ5: Which factors such as class imbalance handling, vocabulary diversity, or schema enrichment contribute most to reliable polymer property extraction?

The experiments showed that vocabulary diversity and schema coverage were more decisive than class

imbalance interventions. Oversampling and class weighting did not yield consistent improvements and were therefore not retained. Instead, the major performance drivers were word coverage (with GloVe outperforming domain-specific Mat2Vec) and the inclusion of character encoders to handle diverse polymer morphologies. Schema enrichment, while introducing rare categories, also added contextual realism to outputs. Overall, reliability was shaped more by embedding coverage and morphological modelling than by explicit balancing techniques.

In sum, the project demonstrated that carefully designed weakly supervised corpora, enriched schemas, and precision-first neural models can deliver robust information extraction for polymers, addressing the central aim set out in Chapter 1.

5.2 Key Findings and Contributions

This dissertation produced three substantive contributions to the field of polymer information extraction.

First, it demonstrated the feasibility of weak supervision for domain corpora. Through the construction of Bulk and Gold corpora, the project showed that weakly labelled full texts can sustain high precision when either scaled across many documents or constrained to a single polymer family. This provides an alternative to expensive manual annotation, showing that corpus design strategies can offset the limitations of noisy labels.

Second, it advanced schema design for scientific NER. By extending beyond the polymer – property – value triad to include METHOD, SOLVENT, and ADDITIVE, the study captured richer contextual information about experimental setups and compositional factors. While performance on these lower-frequency entities was less consistent, their inclusion improved the descriptive completeness of outputs, making extracted records more useful for downstream applications.

Third, it identified BiLSTM+CRF as the most practical model for precision-first extraction. Although transformers such as SciBERT added contextual nuance, the character-aware BiLSTM+CRF achieved comparable or superior performance at far lower computational cost. This model delivered strong boundary integrity and robustness across both large-scale and polymer-specific corpora, establishing it as a reliable backbone for information extraction pipelines in materials science.

Together, these findings show that careful corpus construction, schema enrichment, and targeted neural architectures can significantly reduce the barriers to building structured polymer property datasets.

5.3 Limitations

Despite the contributions of this work, several limitations should be acknowledged.

Corpus quality and scope. The weakly labelled corpora relied on heuristic alignment rules and gazetteers, which introduced noise into entity boundaries and relations. While large-scale weak supervision (Bulk) yielded stable results, annotation sparsity still constrained recall for rare classes such as ADDITIVE. Furthermore, the Gold corpus, although polystyrene-focused and consistent, did not cover the full diversity of polymers encountered in broader literature.

Computational constraints. Experiments were limited to recurrent and SciBERT-based architectures within a restricted GPU budget. This prevented systematic exploration of larger contextual models such as MatSciBERT or PolyBERT, which may have offered stronger domain adaptation but were infeasible within available resources. Fine-tuning SciBERT also required reduced batch sizes and short training schedules, which increased variance across runs.

Schema and evaluation boundaries. The schema was enriched to include METHOD, SOLVENT, and ADDITIVE, but relations between entities (e.g., linking specific methods to measured properties) were not modelled explicitly. Evaluation focused on strict span-level metrics, which are rigorous but do not capture partial correctness in cases such as truncated polymer names. These limitations do not diminish the value of the findings but highlight where further development is required to advance polymer text mining toward more general and fully relational extraction.

5.4 Future Work

The findings of this study highlight several directions for advancing polymer information extraction.

Domain-specialised transformers. Although BiLSTM+CRF proved the most practical backbone, transformer models remain attractive once stability and resource constraints are addressed. Future work could explore pre-trained scientific models such as MatSciBERT or PolyBERT, which incorporate domain vocabulary and may provide stronger contextual embeddings for polymer-specific terms. Careful fine-tuning with larger curated corpora could help realise their potential without the instability observed in SciBERT.

Richer supervision strategies. Weak supervision proved effective for scaling full-text corpora, but hybrid strategies may further improve quality. Active learning pipelines could prioritise uncertain predictions for manual review, while semi-supervised training might exploit unlabelled corpora to regularise models. Combining these approaches would reduce noise and enhance generalisability.

Schema and relational modelling. The current schema captured a broad range of entities but did not model relations explicitly. Future research should extend beyond entity recognition to structured relation extraction, linking polymers to properties, values, and experimental methods. Such capabilities would enable automated population of structured polymer databases and facilitate scientific discovery.

Error-driven refinements. The error analyses identified recurring challenges such as boundary errors in complex polymer names and omissions of numeric tokens. These cases suggest potential for integrating rule-based post-processing or constrained decoding to complement machine learning models, balancing precision with coverage. By addressing these areas, future work can build on the solid foundation established here, moving toward scalable, accurate, and context-rich extraction of polymer property data from scientific text.

This dissertation demonstrated that high-quality information extraction from polymer literature is achievable through carefully designed weakly supervised corpora and precision-first neural architectures. By progressively refining models from CRFs to BiLSTMs and contextual embeddings, it established that character-aware BiLSTM+CRF delivers the most reliable balance of accuracy, interpretability, and scalability for polymer property extraction. The combination of corpus design, schema enrichment, and systematic evaluation answered the central research questions while laying the groundwork for future exploration of domain-specialised transformers and relational modelling. Taken together, the study provides both a practical contribution in the form of robust extraction pipelines and a conceptual contribution by showing how weak supervision, when strategically applied, can unlock value in scientific text at scale.

References

[1] S. Kim *et al.*, "PolyInfo: Polymer database for polymer informatics," *J. Chem. Inf. Model.*, vol. 51, no. 11, pp. 2680–2686, 2011.

[2] A. Mannodi-Kanakkithodi *et al.*, "Scoping polymer informatics: A roadmap for accelerated polymer discovery," *Macromolecules*, vol. 55, no. 2, pp. 373–391, 2022.

[3] S. Park *et al.*, "Text mining metal–organic framework papers," *J. Chem. Inf. Model.*, vol. 58, no. 2, pp. 244–251, 2018.

[4] S. Park *et al.*, "Mining insights on metal–organic framework synthesis from scientific literature texts," *Nat. Commun.*, vol. 13, no. 1, p. 46, 2022.

[5] J. Chen *et al.*, "Automated information extraction from materials science literature," *Patterns*, vol. 3, no. 7, p. 100574, 2022.

[6] Y. Zheng *et al.*, "ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis," *Nat. Mach. Intell.*, vol. 5, no. 11, pp. 1150–1160, 2023.

[7] EPSRC, "Seed Corn project: Polymer informatics for accelerated discovery," Engineering and Physical Sciences Research Council, UK, 2021.

[8] A. Afzal, A. Jha, and R. Ramprasad, "Corpus for polymer named entity recognition and property extraction," *J. Chem. Inf. Model.*, vol. 59, no. 9, pp. 3625–3636, 2019. Extended dataset available at: `https://github.com/Ramprasad-Group/polymer_information_extraction` (accessed Sep. 2025).

[9] P. Shetty and R. Ramprasad, "Machine-guided polymer knowledge extraction using natural language processing: The example of named entity normalization," *J. Chem. Inf. Model.*, vol. 61, no. 11, pp. 5377–5385, 2021.

[10] J. Cheung, T. C. Ling, and K. Lo, "POLYIE: A dataset of information extraction from polymer material scientific literature," *arXiv preprint* arXiv:2311.07715, 2023.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, pp. 282–289, 2001.

[12] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, pp. 1532–1543, 2014.

[13] V. Tshitoyan, J. Dagdelen, L. Weston, *et al.*, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, vol. 571, pp. 95–98, 2019.

[14] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. ACL*, pp. 1064–1074, 2016.

[15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. NAACL*, pp. 260–270, 2016.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[17] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. EMNLP*, pp. 3615–3620, 2019.

[18] A. Gupta, A. Saboo, N. Chetlur, and C. Wolverton, "MatSciBERT: A materials domain language model for text mining and information extraction," *npj Comput. Mater.*, vol. 8, p. 122, 2022.

[19] A. Jain, E. Ward, and R. Ramprasad, "MaterialsBERT: Domain optimized language models for materials science," *arXiv preprint* arXiv:2203.06437, 2022.

The code produced as part of work for this report can be found in the private GitHub repository: https://github.com/07anshul/Polymer-Text-Mining