Reconstructing Manual Information Extraction with *DB-to-Document Backprojection*: Experiments in the Life Science Domain

Mark-Christoph Müller, Sucheta Ghosh, Maja Rey, Ulrike Wittig, Wolfgang Müller and Michael Strube
Heidelberg Institute for Theoretical Studies gGmbH, Heidelberg, Germany
{mark-christoph.mueller,sucheta.ghosh,maja.rey,
ulrike.wittig,wolfgang.mueller,michael.strube}@h-its.org

Abstract

We introduce a novel scientific document processing task for making previously inaccessible information in printed paper documents available to automatic processing. We describe our data set of scanned documents and data records from the biological database SABIO-RK, provide a definition of the task, and report findings from preliminary experiments. Rigorous evaluation proved challenging due to lack of gold-standard data and a difficult notion of correctness. Qualitative inspection of results, however, showed the feasibility and usefulness of the task.

1 Introduction

Research results from the life sciences are mainly published in the form of written journal or conference papers, even though these results often take the form of measurements of experimental parameters, which would more appropriately be stored in a structured, machine-readable form. While there is some tendency towards directly publishing experimental data, e.g. on SourceData (Liechti et al., 2016) or (for environmental data) PANGAEA[1], this is not the norm yet, and does not help with the huge body of data already published in the conventional literature. It is common practice in the life sciences, therefore, to manually extract information (including measurements and the experimental conditions underlying them) from natural language documents, and to use it to populate biological databases. This process is called *biocuration* (International Society for Biocuration, 2018) and comprises, for every document, 1) identification and mark-up of curatable information, 2) data extraction, normalization, and consolidation, and 3) database insertion. Despite constant improvements in NLP technology, biocuration involves significant human labor (mostly reading) (Oughtred et al., 2019; Huang et al., 2020; Wu et al., 2020; Abdelhakim et al., 2020), because data *quality* (i.e. correctness and integrity) has priority over *quantity* (i.e. more quickly available, but potentially less reliable, data), and the error rates of current NLP systems are still considered too high (Karp, 2016). For reasons of ergonomics and ease of handling (Buchanan and Loizides, 2007; Köpper et al., 2016; Clinton, 2019), the identification and mark-up step often involves paper printouts and highlighter pens,[2] like in the example page in Figure 1.

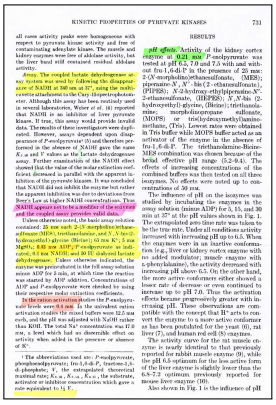


Figure 1: Page with mark-up (best viewed in color).

As mere intermediate products of the curation process, the manually highlighted printouts are only required until all data from the respective document has been curated, and they will normally be archived afterwards. We argue, however, that the printouts contain even *more* information which curation simply does not make full use of: First, some document sections, although containing highlighting, will *not* lead to the creation of a record in the biological database (see our results in Section 5.3). Yet, this highlighting can still be regarded as a kind of relevance annotation, produced by life science

[1] www.pangaea.de

[2] This is true for our own group, and has been corroborated in 2016 by an informal, unpublished survey among 21 curators from 15+ biological databases. The survey showed that a considerable number of curators rely on paper printouts for close reading and / or highlighting of important information.

Proceedings of the First Workshop on Scholarly Document Processing, pages 81–90
Online, November 19, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

domain experts through attentive, task-oriented reading. Obviously, this information should be useful, e.g. for the analysis of how important information is dispersed over a scientific document. Second, for those database records that *are* created from highlighted document sections, the reference to that section is normally not preserved. Again, an obvious way to use this information is to allow users of the biological database to visually trace the record to its source in the document, including the original context.

In this paper, we describe our approach towards re-purposing scientific document printouts which were manually highlighted during biocuration. More precisely, our research question is: *Given records of curated information from the database and the original, scanned source document, (to what degree) can we recover the document* section *that a particular record was extracted from?* We consider this to be a novel scientific document processing task, and propose to refer to it as **DB-to-document backprojection**. The remainder of the paper is structured as follows. In Section 2 we describe the data basis of our work. Section 3 introduces the *highlighted text extraction* task, which we consider as self-contained and only loosely linked to the main task. Section 4 deals with the actual DB-to-document backprojection task, provides a precise definition, and describes our processing steps. Section 5 presents some preliminary experiments, results, and error analysis.Initially, this section will also discuss our approach to evaluation. In Section 6 we discuss some related work, and Section 7 contains our conclusions and directions for the future. Note that, although our data is from the life sciences, the task is relevant for all domains where manual information extraction is performed on natural language documents (like e.g. in Lipani et al. (2014), where information is extracted from IR research papers in the form of machine-readable 'nanopublications').

2 Data

The work in this paper is based on two related data sets, which have been collected in the SABIO-RK Biochemical Reaction Kinetics Database project[3]. SABIO-RK is a curated database containing structured information about biochemical reactions and their corresponding kinetics (Wittig et al., 2017, 2018). The **document data set** is an electronic version of our archive of $6,000$+ manually highlighted printouts of documents from the life science domain, which have been curated in the 10+ years of our database's existence. Over the years, numerous different curators were involved in the manual mark-up. Different highlighter colors were used, sometimes even within the same document (see Figure 1). In case of equivalent information appearing repeatedly in the same document, curators generally attempted to be economical and to avoid redundancy by highlighting only the *most appropriate* appearance, which is often, but not always, the *first* appearance. While the mark-up was performed in a completely unrestricted manner (cf. below), in the vast majority of cases, highlighting was applied directly to words or lines (cf. Figure 1), which greatly helped in extracting the highlighted text (cf. Section 3). In some rare cases, curators selected whole sections by drawing a vertical line at the section's margin. Also, data in tables was sometimes highlighted on the cell level, while in other cases, only the column header, the table header, or even the table caption was highlighted. We created an electronic version of the document collection by scanning and OCR-processing all papers[4], which resulted in a sandwich PDF for each document with the (partially highlighted) background superimposed with the extracted text. OCR was performed with commercial software (Alaris Capture Pro), which was used out-of-the-box. The total number of tokens in the 98 documents is $630,153$, with $6,430$ tokens/document on average.

The second data set is the **record data set** which contains measurements of kinetic parameters that were extracted from individual documents from the document collection in the course of manual curation. Each of the $2,916$ records in this data set is linked to exactly one source document (via its PubMed ID), but no lower-level links (to pages or lines) exist. Each document, in turn, can be linked to an arbitrary number of records (29.76 records/document on average). It has to be noted that the above count of $2,916$ records contains a considerable number of *multiple counts*. This is true in particular for records of type *experimental condition* (cf. below), and is due to the fact that often, several measurements are performed under identical experimental conditions. For scoring and evaluation, however, this does not make a differ-

[3] http://sabio.h-its.org/

[4] For the experiments reported in this paper, we only use a subset of 98 documents.

ence, because we conflate semantically identical records before analysis. There are two main types of records, *experimental condition* and *parameter*. Each record consists of three to six attribute-value (a-v) pairs. Figures 2 and 3 show one example of each type of record.

```
conditionName: 'pH',
startValue   :  7.7,
buffer       : '0.10 M Tris-HCl,
               100 mM KCl, 1 mM DTT,
               4.0 mM MgCl2,
               10% Glycerol'
```

Figure 2: Record of type *experimental condition* with three a-v pairs featuring one numeric, one atomic string, and one complex string value.

```
parameterName     : 'Km',
unitName          : 'μM',
startValue        :  123,
standardDeviation:  12,
associatedSpecies: 'Acetyl-CoA'
```

Figure 3: Record of type *parameter* with five a-v pairs, featuring two numeric and three atomic string values.

Note that we only consider a subset of all a-v pairs available for each record: Some attributes have unspecific values (e.g. `role:'Variable'`) which are not useful for searching. Also, most attributes have a variant with a *normalized* value, which does not appear in the text. With the exception of the experimental conditions' `buffer` attribute, all values are atomic. Therefore, the `buffer` attribute will be handled differently in the second phase of backprojection (see Section 4.2).

3 Highlighted Text Extraction

Highlighted text extraction comprises 1) extracting from each sandwich PDF both the searchable plain text and a background image for every page, 2) detecting highlighted areas in the background images, and 3) mapping the detected image areas to the extracted text. The workflow is shown in Figure 4. We use both `pdftohtml` and `pdftotext` from the Poppler[5] library to extract data from the scanned and OCR-processed sandwich PDF documents from our collection. The only task of `pdftohtml` is to extract, from each page, a PNG image with the non-textual background, which also includes the color-marked areas. These images were already generated during OCR processing and consist of document pages from which pixels that were detected as belonging to text were removed by inpainting (see 'Page background image' in the lower left part of Figure 4). `pdftotext`, on the other hand, is used to extract the text that was previously recognized by OCR. It produces one XML file for the document, incl. bounding boxes on the token-level. These tokens reflect the original document layout, but come in correct reading order even for multi-column documents. The second step makes use of some simple image processing. As described above (Section 2), document highlighting can come in any color, so searching for areas of any *particular* color (like e.g. yellow) is not an option. Instead, our algorithm combines the facts that 1) highlighting is always *non-grey* and 2) shades of grey in the RGB color model are characterized by identical, or at least highly similar, values in the R, G, and B components.[6] We create a binarized version of each page by going over all pixels in a copy of the original image and setting each pixel to 'black' if the difference between the R, G, and B components is above a threshold of 50 (i.e. if the pixel is non-grey), and setting it to 'white' otherwise. The resulting image, then, contains regions with higher and lower density of black pixels (see 'Binarized page background image' in Figure 4). In the last step, text tokens are labelled as highlighted if their bounding boxes (from the XML file), when projected to the binarized image, cover an area that is at least 50% black. While this process is very simple, we found it to work surprisingly well, at least for the very frequent cases where the highlighting was applied directly to words or lines, yielding almost perfect extraction accuracy on most of the images we inspected. Of the $630,153$ tokens in our data set, only $39,071$ (6.2%) were detected as containing highlighting.

4 DB-to-Document Backprojection

4.1 Task Definition

DB-to-document backprojection attempts to reconstruct the manual information extraction performed during database curation, by recovering those document sections that the curated database records were extracted from. It works by matching database record values (as strings) to plain text from document sections. More precisely, we define the task as follows: Let **D** be the document data set, **R** the record data set, **R(d)** the set of

[5] https://poppler.freedesktop.org

[6] https://en.wikipedia.org/wiki/Grey

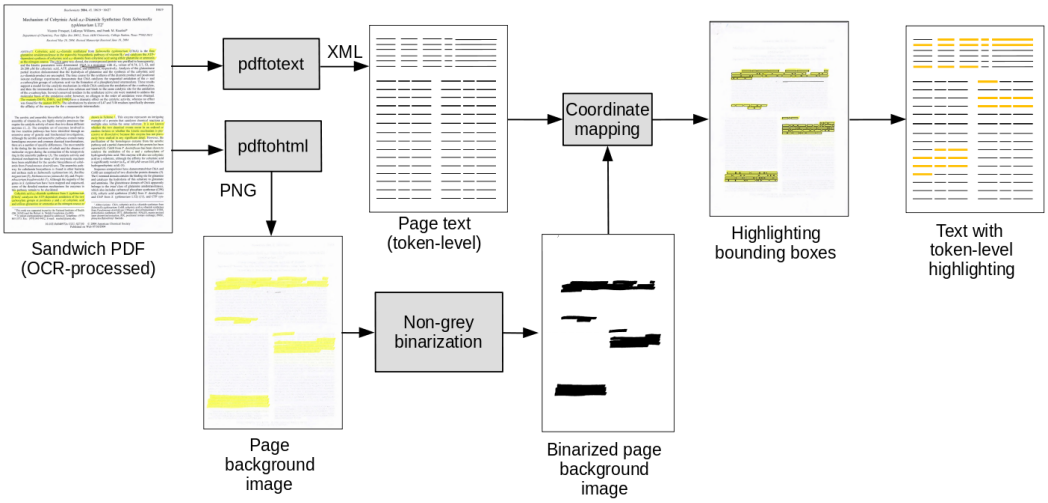


Figure 4: Highlighted text extraction workflow (best viewed in color).

records that were extracted from document **d** $\in$ **D**, and **V(r)** the set of values belonging to record **r** $\in$ **R(d)**. Also, let **SEC(d, sec_size)** be the set of document sections of *sec_size* tokens into which document **d** $\in$ **D** can be segmented. Then, for every document **d** $\in$ **D**, for every section **s** $\in$ **SEC(d, sec_size)**, and for every record **r** $\in$ **R(d)**, a backprojection score between 0.0 and 1.0 is computed by counting how many of the values in **V(r)** can be matched to the tokens in **s**, and normalizing by the total number of values in **V(r)**. The result is a list of $< record, section, score >$ tuples for every document, from which the most *plausible* backprojections still needs to be selected (cf. Section 5.1).

The following is worth noting. First and foremost, the above definition reflects the fact that there is no simple notion of a *correct* backprojection of a database record to a document section, neither in our data sets nor, arguably, in reality. In part, this is because the same (or highly similar) information can appear in more than one section of a document. Second, the value of *sec_size* is important because, by specifying the number of tokens that are considered at the same time, it might penalize records with a comparably large number of values. At the same time, however, an excessively large *sec_size* will undermine the whole endeavour because it will be difficult to locate the actual matched values within the section. Also, with increasing *sec_size*, there is a growing risk of clustering values which are actually completely unrelated, creating spurious backprojection results. Finally, the role of automatically detected highlighting for DB-to-document backprojection is still unclear. Since obtaining this highlighting information was the prime reason for scanning the paper printouts in the first place, a rather strong contribution of this feature is desirable. One obvious role of highlighting is that of a filter for preventing *non*-highlighted tokens from being potential backprojection targets.

4.2 Processing Steps

The following two steps are performed for every document **d** $\in$ **D** and for every record **r** $\in$ **R(d)**. In the first step, **search term creation**, the non-empty values in **V(r)** are converted into search terms. Initially, one search term list is created for each non-empty **v** $\in$ **V(r)**. Thus, a record with three values will yield as many search term lists with *one* term each (an example is given below). These three lists are *complementary*, i.e. we try to match elements of as many of them as possible in a given document section. In order to improve matching and to capture variations introduced by spelling alternatives and / or OCR errors, we apply the following heuristics, which add *alternative* search terms to individual search term lists: For numerical values with a decimal point (e.g. '9.5'), we add a term where that character is replaced by a comma. If OCR_OPTIMIZE=TRUE: For string values containing the μ character (e.g. 'μg/ml' or 'μM'), we add a term where that character is replaced by a 'p', which is a common OCR error / substitution. Likewise, for string values containing a lowercase 'm' character at the end (e.g. Km), we add several terms where that character is replaced by 'tn', 'ni', and a combination of commas, which are common OCR errors if the 'm' appears as a subscript. If

USE_SYNONYMS=TRUE: For string values representing chemical compound names, we consult a look-up table and add synonyms, spelling variants, or abbreviations as alternative terms.

For illustration, with OCR_OPTIMIZE=TRUE, the record of type *parameter* from Figure 3 above yields the following list of search term lists, with a range of possible matches from zero to five, corresponding to its number of values. Note the spelling variants for the first and second value. During term matching (cf. below), only *one* item per search term list needs to match in order for the value (first item in each list) to match.

```
[['Km', 'Ktn', 'Kni', 'K,,,'],
 ['μM', 'pM'],
 ['123'],
 ['12'],
 ['Acetyl-CoA']]
```

As mentioned in Section 2, the `buffer` attribute of records of type *experimental condition* is special because its value is a manually edited, comma-separated string containing several chemical substance names (see Figure 2). We split each value string into a list of individual substance names, and add each of these names as an *additional* search term list for that record.

Then, in the second step, the actual **term matching** is performed in the following way: For each document $\mathbf{d} \in \mathbf{D}$, we iterate over all tokens in **d** (extracted from the XML output of `pdftotext`, cf. Section 3), all records $\mathbf{r} \in \mathbf{R(d)}$, and all search term lists created for the respective **r** in the previous step. Then, we iterate over the terms in each search term list, trying to match each one in turn. Matching is done simply by using regular expressions. If a term can be matched to a token, we collect the matching record's ID and the matched value in the token's matchlist, and move on to the next search term list. This matching process is performed only once, and it is the same regardless of the value of *sec_size*.

Next, sections of different sizes are created by moving a window of size *sec_size* over all tokens in **d**, one token at a time. These sections are the potential targets for backprojection. In our experiments, *sec_size* ranges from 3 to 39, in steps of 3, and the following steps are performed for each value of *sec_size*. If the first token in a potential section has a non-empty match list, a matching result for the entire section is computed in the following way: First, all record IDs with a match anywhere in the section are collected. Then, for each of these records, a section score is computed by counting the *distinct* matches in the section and normalizing that with the maximum number of possible matches. The restriction to *distinct* values means that if a term matches more than one token in a section, it is only counted once *for each record*. Without this restriction, values appearing repeatedly in the same section (like e.g. unit names) would incorrectly boost the scores for the respective records. In most cases, a record will match several sections with different scores, but we only select the top scoring sections for each record. In the end, this results in a mapping of record IDs to the top score for these records and a list of sections with this score.

In addition, we introduce the following experimental parameters into the backprojection step: **HL_ROLE**: If HL_ROLE=IGNORE, highlighting information is not used, if HL_ROLE=ONLY, only highlighted tokens (as determined by highlighted text extraction (Section 3) will be considered for matching. **MIN_MATCHES**: The minimum number of values for a record that need to match in a section in order for that section to be considered. MIN_MATCHES < 2 will yield a lot of spurious matches. **REQUIRE_NUM_MATCH**: If REQUIRE_NUM_MATCH=TRUE, at least one of the matched record values in a section must be numeric. This is based on the rationale that numeric values are more distinctive than e.g. matches for parameter or unit names.

5 Experiments

5.1 A Note on Evaluation

As described in the task definition in the previous Section 4, the result of performing a DB-to-document backprojection run with a given set of parameters on a single document is a mapping of each record in the document to those section(s) that yielded the maximum score for that record (possibly none). While the *inspection* of this result (including visualisation, cf. below) is straightforward, an actual quantitative *evaluation* is more difficult. This evaluation would have to include the identification of true and false positives (i.e. records that are backprojected to correct resp. incorrect document sections) and false negatives (i.e. records that were *not* backprojected even though a document section with a sufficient fraction of the record's values exists). This form of evaluation is out of the scope of the present paper. The most obvious reason is that, at least at present, no annotated gold-level data is

available which specifies, for each record, one or more document sections as the correct backprojection target. In addition, it will become clear in what follows that there not even is a simple notion of a *correct* backprojection.

5.2 Preliminary Experiments

We performed a couple of preliminary experiments, at first setting the parameters to HL_ROLE= IGNORE, REQUIRE_NUM_MATCH= TRUE, and MIN_MATCHES= 2. On the level of the individual document, inspection of experimental results is straightforward: Figure 5 contains a heatmap with the result for one document which shows, for each record[7] (rows) and different values of *sec_size* (columns), the maximum score (top of cell) and the number of sections with this score (bottom of cell). Cell values are only displayed if they change from column to column. The row headers contain the ID and the total number of values for each record (i.e. the size of **V(r)**), which is the maximum number of possible matches.

Figure 5 allows to make several observations: First, two records (269787 and 269763) could not be backprojected at all under the applied settings, which is visible in their score being 0.000 throughout the whole range of *sec_size* values. The overall highest score of 0.833 was reached by six records, each of which has six potentially matchable values, in precisely one section. However, for the first, third, and fifth record, the best match was found for *sec_size*=9, while for the fourth and sixth one, *sec_size* had to be as high as 24, and even 30 for the second one. In other words, while the six values of some of the records were found in close proximity to each other, for others, they were scattered over a range of more than twice resp. three times that size.

Next, we inspect the effect of *one possible way* of using automatically detected highlighting information, by re-running the previous experiment with HL_ROLE= ONLY, i.e. we require the presence of highlighting for a token to be part of a match. Ideally, this should improve backprojection *precision*, by eliminating spurious matches. Given the low incidence of highlighting in our data (only 6.2% of all tokens, cf. Section 3), this might drastically reduce the number of records that can be matched at all. What is more, given the unconstrained way in which the highlighting was applied by the curators, care has to be taken that the presence (and, more importantly, the *absence*) of highlighting is not over-interpreted.

Figure 6 displays the result for the same document with HL_ROLE= ONLY. Some effects are clearly visible: Two records are no longer backprojected at all.[8] For two other records (163378193 and 269766), the maximum scores are reduced (from 0.833 to 0.500 and 0.333, respectively).

In summary, the above discussion shows that the heatmap visualisation provides a reasonable and reasonably compact representation of a complete DB-to-document backprojection result. It allows to identify record-to-section mappings with varying plausibility, on the basis of how widely scattered the values are in the target sections. This makes it useful for the comparison of different results, like the two results with HL_ROLE= IGNORE and ONLY. The actual verification and *qualitative* evaluation and error analysis, however, requires a more detailed approach (cf. Section 5.3).

5.3 Detailed Analysis

For detailed (error) analysis, records can visually be 'projected' to their automatically detected target sections. Figure 7 shows this for one page each from two documents. These results were created with OCR_OPTIMIZE and USE_SYNONYMS= TRUE, and HL_ROLE= IGNORE. Boxes on the right-hand side show (at the top) the record ID and the matching *sec_size* and score, followed by the section text as recognized by OCR, and all search terms, one search term list (cf. Section 4.2) per row. Unmatched values are given in bold red. The following points are interesting to note. For the first record on the top page, the system failed to identify the 'NaCL' token, which was caused by an OCR error which misread 'NaCL' as 'NaCI'. The bottom three records on the top page exemplify the positive effect of the OCR_OPTIMIZATION, since 'μM' was only matched because of the replacement 'pM' (the same is true for several records in the lower page). It is also instructive to see that, by setting HL_ROLE= IGNORE, two matches could be found in sections without any highlighting. In the lower page, we see the positive effect of USE_SYNONYMS= TRUE in the third and forth record, where the replacement 'NAD' was found

[7] For semantically identical records, only one, arbitrarily selected ID is provided, because all other records have exactly the same result.

[8] Both are actually false negatives, which were not highlighted but appeared as part of a table, which was highlighted on the title level.

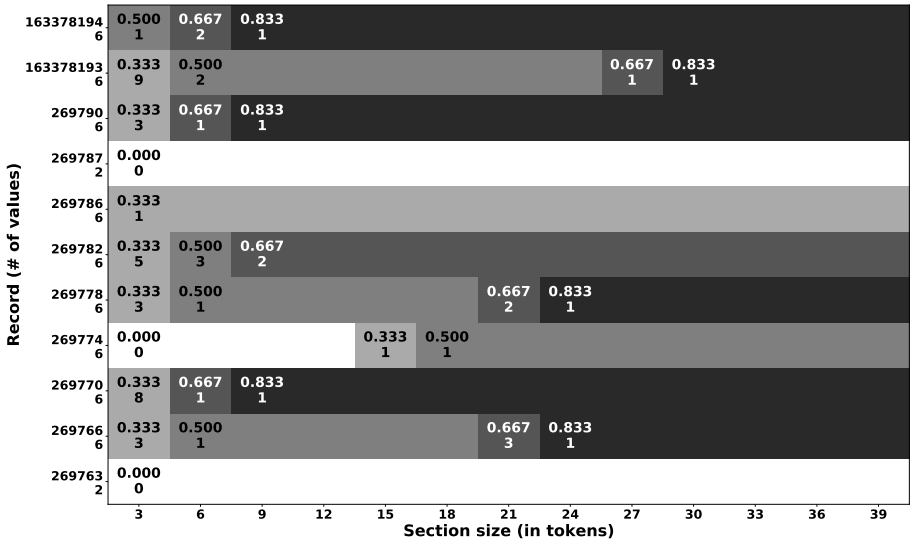


Figure 5: Single Document-level result, HL_ROLE=IGNORE

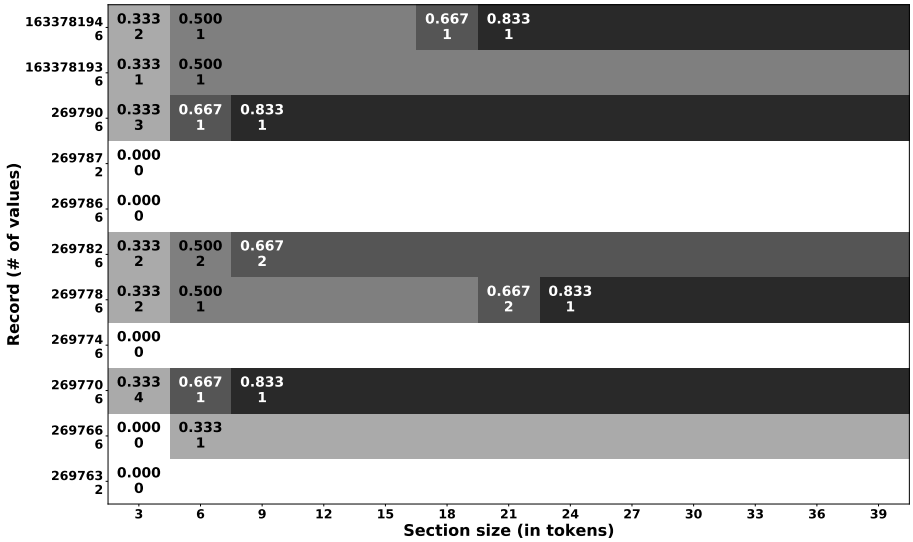


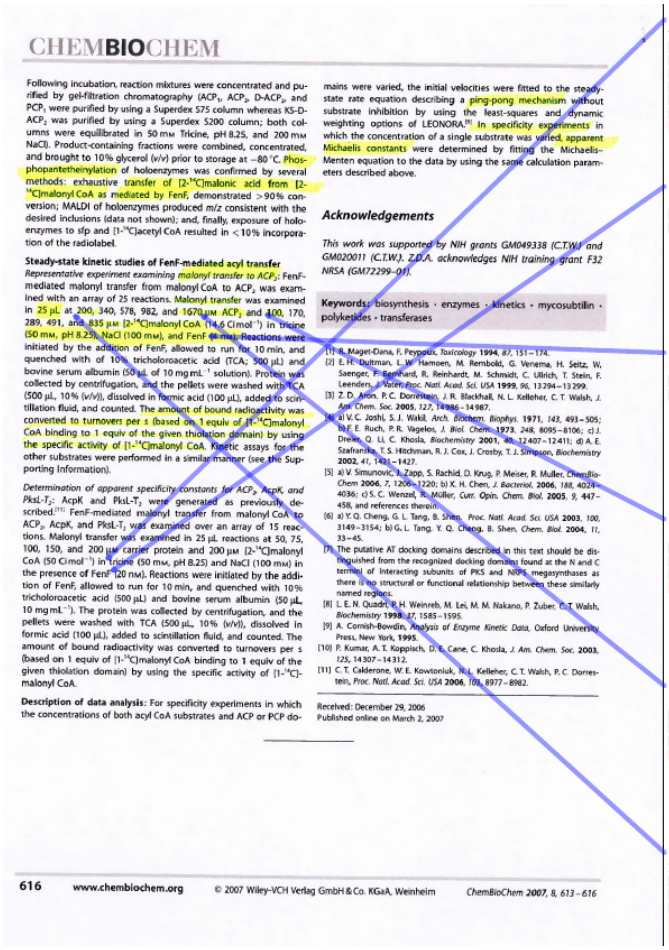
Figure 6: Single Document-level result, HL_ROLE=ONLY

instead of the originally required 'NAD+'. Finally, the lower page also shows that highlighting is not necessarily associated with extracted information.

6 Related Work

DB-to-document backprojection is related to several NLP and document processing tasks, but it is quite special in that it combines 1) OCR processing of scanned documents, 2) information extraction, 3) template matching, and 4) strictly string-based (as opposed to semantic) matching. Scanned paper documents are much less often subject of text or information extraction than born-digital documents like PDFs. *Robust reading*[9] is a common term under which several approaches are collected. A recent approach in this area is DeepReader (Vishwanath et al., 2018), which is a document understanding approach which seamlessly integrates low-level OCR with recognition of higher-level document structure and, to a certain extent, content. *Document Visual Question Answering* (Mathew et al., 2020), on the other hand, analyses scanned documents beyond mere OCR of text content, including manually applied highlighting, for answering questions about the documents' content. On the other hand, information extraction and semantic representation or modelling, especially from publications from the bio domain, is a very active field (Vahdati et al., 2019; Anteghini et al., 2020).

[9] https://rrc.cvc.uab.es/



CHEMBIOCHEM

Following incubation, reaction mixtures were concentrated and purified by gel-filtration chromatography (ACP_1, ACP_2, D-ACP_2, and PCP_2 were purified by using a Superdex 575 column whereas KS-D-ACP_2 was purified by using a Superdex S200 column; both columns were equilibrated in 50 mM Tricine, pH 8.25, and 200 mM NaCl). Product-containing fractions were combined, concentrated, and brought to 10% glycerol (v/v) prior to storage at −80 °C. Phosphopantetheinylation of holoenzymes was confirmed by several methods: exhaustive transfer of [2-^{14}C]malonic acid from [2-^{14}C]malonyl CoA as mediated by FenF, demonstrated >90% conversion; MALDI of holoenzymes produced *m/z* consistent with the desired inclusions (data not shown); and, finally, exposure of holoenzymes to *sfp* and [1-^{14}C]acetyl CoA resulted in <10% incorporation of the radiolabel.

Steady-state kinetic studies of FenF-mediated acyl transfer

Representative experiment examining malonyl transfer to ACP_2: FenF-mediated malonyl transfer from malonyl CoA to ACP_2 was examined with an array of 25 reactions. Malonyl transfer was examined in 25 μL at 200, 340, 578, 982, and 1670 μM ACP_2 and 100, 170, 289, 491, and 835 μM [2-^{14}C]malonyl CoA (14.6 C mol^{-1}) in tricine (50 mM, pH 8.25), NaCl (100 mM), and FenF (4 nM). Reactions were initiated by the addition of FenF, allowed to run for 10 min, and quenched with of 10% tricholoroacetic acid (TCA; 500 μL) and bovine serum albumin (50 μL of 10 mg mL^{-1} solution). Protein was collected by centrifugation, and the pellets were washed with TCA (500 μL, 10% (v/v)), dissolved in formic acid (100 μL), added to scintillation fluid, and counted. The amount of bound radioactivity was converted to turnovers per s (based on 1 equiv of [1-^{14}C]malonyl CoA binding to 1 equiv of the given thiolation domain) by using the specific activity of [1-^{14}C]malonyl CoA. Kinetic assays for the other substrates were performed in a similar manner (see the Supporting Information).

Determination of apparent specificity constants for ACP_2, AcpK, and $PksL\text{-}T_2$: AcpK and $PksL\text{-}T_2$ were generated as previously described.[11] FenF-mediated malonyl transfer from malonyl CoA to ACP_2, AcpK, and $PksL\text{-}T_2$ was examined over an array of 15 reactions. Malonyl transfer was examined in 25 μL reactions at 50, 75, 100, 150, and 200 μM carrier protein and 200 μM [2-^{14}C]malonyl CoA (50 C mol^{-1}) in tricine (50 mM, pH 8.25) and NaCl (100 mM) in the presence of FenF (20 nM). Reactions were initiated by the addition of FenF, allowed to run for 10 min, and quenched with 10% tricholoroacetic acid (500 μL) and bovine serum albumin (50 μL, 10 mg mL^{-1}). The protein was collected by centrifugation, and the pellets were washed with TCA (500 μL, 10% (v/v)), dissolved in formic acid (100 μL), added to scintillation fluid, and counted. The amount of bound radioactivity was converted to turnovers per s (based on 1 equiv of [1-^{14}C]malonyl CoA binding to 1 equiv of the given thiolation domain) by using the specific activity of [1-^{14}C]-malonyl CoA.

Description of data analysis: For specificity experiments in which the concentrations of both acyl CoA substrates and ACP or PCP domains were varied, the initial velocities were fitted to the steady-state rate equation describing a ping-pong mechanism without substrate inhibition by using the least-squares and dynamic weighting options of LEONORA.[9] In specificity experiments in which the concentration of a single substrate was varied, apparent Michaelis constants were determined by fitting the Michaelis–Menten equation to the data by using the same calculation parameters described above.

Acknowledgements

This work was supported by NIH grants GM049338 (C.T.W.) and GM020011 (C.T.W.). Z.D.A. acknowledges NIH training grant F32 NRSA (GM72299–01).

Keywords: biosynthesis · enzymes · kinetics · mycosubtilin · polyketides · transferases

[1] B. Maget-Dana, F. Peypoux, *Toxicology* **1994**, *87*, 151–174.
[2] E. H. Duitman, L. W. Hamoen, M. Rembold, G. Venema, H. Seitz, W. Saenger, F. Bernhard, R. Reinhardt, M. Schmidt, C. Ullrich, T. Stein, F. Leenders, J. Vater, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 13294–13299.
[3] Z. D. Aron, P. C. Dorrestein, J. R. Blackhall, N. L. Kelleher, C. T. Walsh, *J. Am. Chem. Soc.* **2005**, *127*, 14986–14987.
[4] a) V. C. Joshi, S. J. Wakil, *Arch. Biochem. Biophys.* **1971**, *143*, 493–505; b) F. E. Ruch, P. R. Vagelos, *J. Biol. Chem.* **1973**, *248*, 8095–8106; c) J. Dreier, Q. Li, C. Khosla, *Biochemistry* **2001**, *40*, 12407–12411; d) A. E. Stafansson, T. S. Hitchman, R. J. Cox, J. Crosby, T. J. Simpson, *Biochemistry* **2002**, *41*, 1421–1427.
[5] a) V. Simunovic, J. Zapp, S. Rachid, D. Krug, P. Meiser, R. Muller, *ChemBioChem* **2006**, *7*, 1206–1220; b) X. H. Chen, *J. Bacteriol.* **2006**, *188*, 4024–4036; c) S. C. Wenzel, R. Muller, *Curr. Opin. Chem. Biol.* **2005**, *9*, 447–458, and references therein.
[6] a) Y. Q. Cheng, G. L. Tang, B. Shen, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 3149–3154; b) G. L. Tang, Y. Q. Cheng, B. Shen, *Chem. Biol.* **2004**, *11*, 33–45.
[7] The putative AT docking domains described in this text should be distinguished from the recognized docking domains found at the N and C termini of interacting subunits of PKS and NRPS megasynthases as there is no structural or functional relationship between these similarly named regions.
[8] L. E. N. Quadri, R. H. Weinreb, M. Lei, M. M. Nakano, P. Zuber, C. T. Walsh, *Biochemistry* **1998**, *37*, 1585–1595.
[9] A. Cornish-Bowden, *Analysis of Enzyme Kinetic Data*, Oxford University Press, New York, **1995**.
[10] P. Kumar, A. T. Koppisch, D. E. Cane, C. Khosla, *J. Am. Chem. Soc.* **2003**, *125*, 14307–14312.
[11] C. T. Calderone, W. E. Kowtoniuk, N. L. Kelleher, C. T. Walsh, P. C. Dorrestein, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8977–8982.

Received: December 29, 2006
Published online on March 2, 2007

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

ID: 163416093 Max. score 0.857 @ sec size 9
tricine (50 mM, pH 8.25) and NaCl (100 mM)
['pH'] --> pH
['8.25', '8.25'] --> 8.25
['50', '50.0', '50,0'] --> 50
['mM'] --> mM
['tricine'] --> tricine
['100', '100.0', '100,0'] --> 100
['NaCl']

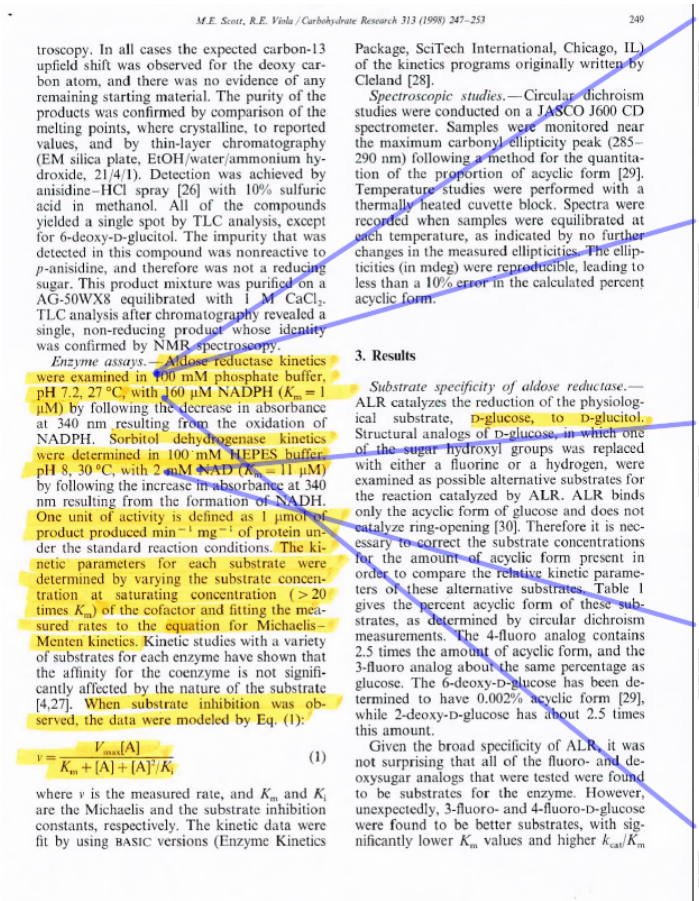
ID: 271756 Max. score 1.000 @ sec size 3
(20 nM). Reactions
['nM'] --> nM
['20', '20.0', '20,0'] --> 20

ID: 271719 Max. score 1.000 @ sec size 3
(4 nM). Reactions
['nM'] --> nM
['4', '4.0', '4,0'] --> 4

ID: 271716 Max. score 0.750 @ sec size 9
100, 170, 289, 491, and 835 pM [2-,4C]malonyl CoA
['μM', 'pM'] --> pM
['100', '100.0', '100,0'] --> 100
['835', '835.0', '835,0'] --> 835
['Malonyl-CoA']

ID: 271716 Max. score 0.750 @ sec size 9
pM ACP2 and 100, 170, 289, 491, and 835
['μM', 'pM'] --> pM
['100', '100.0', '100,0'] --> 100
['835', '835.0', '835,0'] --> 835
['Malonyl-CoA']

ID: 271715 Max. score 0.600 @ sec size 9
200, 340, 578, 982, and 1670 pM ACP2 and
['μM', 'pM'] --> pM
['200', '200.0', '200,0'] --> 200
['1670', '1670.0', '1670,0'] --> 1670
['[Acyl-carrier']
['protein]']



M.E. Scott, R.E. Viola / Carbohydrate Research 313 (1998) 247–253 249

troscopy. In all cases the expected carbon-13 upfield shift was observed for the deoxy carbon atom, and there was no evidence of any remaining starting material. The purity of the products was confirmed by comparison of the melting points, where crystalline, to reported values, and by thin-layer chromatography (EM silica plate, EtOH/water/ammonium hydroxide, 21/4/1). Detection was achieved by anisidine–HCl spray [26] with 10% sulfuric acid in methanol. All of the compounds yielded a single spot by TLC analysis, except for 6-deoxy-D-glucitol. The impurity that was detected in this compound was nonreactive to *p*-anisidine, and therefore was not a reducing sugar. This product mixture was purified on a AG-50WX8 equilibrated with 1 M $CaCl_2$. TLC analysis after chromatography revealed a single, non-reducing product whose identity was confirmed by NMR spectroscopy.

Enzyme assays.—Aldose reductase kinetics were examined in 100 mM phosphate buffer, pH 7.2, 27 °C, with 160 μM NADPH (K_m = 1 μM) by following the decrease in absorbance at 340 nm resulting from the oxidation of NADPH. Sorbitol dehydrogenase kinetics were determined in 100 mM HEPES buffer, pH 8, 30 °C, with 2 mM NAD (K_m = 11 μM) by following the increase in absorbance at 340 nm resulting from the formation of NADH. One unit of activity is defined as 1 μmol of product produced min^{-1} mg^{-1} of protein under the standard reaction conditions. The kinetic parameters for each substrate were determined by varying the substrate concentration at saturating concentration (>20 times K_m) of the cofactor and fitting the measured rates to the equation for Michaelis–Menten kinetics. Kinetic studies with a variety of substrates for each enzyme have shown that the affinity for the coenzyme is not significantly affected by the nature of the substrate [4,27]. When substrate inhibition was observed, the data were modeled by Eq. (1):

$$v = \frac{V_{max}[A]}{K_m + [A] + [A]^2/K_i} \quad (1)$$

where v is the measured rate, and K_m and K_i are the Michaelis and the substrate inhibition constants, respectively. The kinetic data were fit by using BASIC versions (Enzyme Kinetics Package, SciTech International, Chicago, IL) of the kinetics programs originally written by Cleland [28].

Spectroscopic studies.—Circular dichroism studies were conducted on a JASCO J600 CD spectrometer. Samples were monitored near the maximum carbonyl ellipticity peak (285–290 nm) following a method for the quantitation of the proportion of acyclic form [29]. Temperature studies were performed with a thermally heated cuvette block. Spectra were recorded when samples were equilibrated at each temperature, as indicated by no further changes in the measured ellipticities. The ellipticities (in mdeg) were reproducible, leading to less than a 10% error in the calculated percent acyclic form.

3. Results

Substrate specificity of aldose reductase.—ALR catalyzes the reduction of the physiological substrate, D-glucose, to D-glucitol. Structural analogs of D-glucose, in which one of the sugar hydroxyl groups was replaced with either a fluorine or a hydrogen, were examined as possible alternative substrates for the reaction catalyzed by ALR. ALR binds only the acyclic form of glucose and does not catalyze ring-opening [30]. Therefore it is necessary to correct the substrate concentrations for the amount of acyclic form present in order to compare the relative kinetic parameters of these alternative substrates. Table 1 gives the percent acyclic form of these substrates, as determined by circular dichroism measurements. The 4-fluoro analog contains 2.5 times the amount of acyclic form, and the 3-fluoro analog about the same percentage as glucose. The 6-deoxy-D-glucose has been determined to have 0.002% acyclic form [29], while 2-deoxy-D-glucose has about 2.5 times this amount.

Given the broad specificity of ALR, it was not surprising that all of the fluoro- and deoxysugar analogs that were tested were found to be substrates for the enzyme. However, unexpectedly, 3-fluoro- and 4-fluoro-D-glucose were found to be better substrates, with significantly lower K_m values and higher k_{cat}/K_m

ID: 159982103 Max. score 1.000 @ sec size 6
100 mM phosphate buffer, pH 7.2,
['pH'] --> pH
['7.2', '7,2'] --> 7.2
['100', '100.0', '100,0'] --> 100
['mM'] --> mM
['phosphate'] --> phosphate

ID: 159982102 Max. score 1.000 @ sec size 9
100 mM phosphate buffer, pH 7.2, 27 °C, with
['27', '27.0', '27,0'] --> 27
['°C'] --> °C
['100', '100.0', '100,0'] --> 100
['mM'] --> mM
['phosphate'] --> phosphate

ID: 101200 Max. score 1.000 @ sec size 6
NAD (Km = 11 pM) by
['μM', 'pM'] --> pM
['Km', 'K$_m$', 'K$_M$', 'K$_{,m}$', 'K$_{m}$', 'K$_{,,}$', 'Ktn', 'Kni'] --> Km
['11', '11.0', '11,0'] --> 11
['NAD+', 'NAD', 'DPN', 'Nadide', 'beta-NAD+'] --> NAD

ID: 101200 Max. score 1.000 @ sec size 6
mM NAD (Km = 11 pM)
['μM', 'pM'] --> pM
['Km', 'K$_m$', 'K$_M$', 'K$_{,m}$', 'K$_{m}$', 'K$_{,,}$', 'Ktn', 'Kni'] --> Km
['11', '11.0', '11,0'] --> 11
['NAD+', 'NAD', 'DPN', 'Nadide', 'beta-NAD+'] --> NAD

ID: 101105 Max. score 1.000 @ sec size 3
160 pM NADPH
['μM', 'pM'] --> pM
['160', '160.0', '160,0'] --> 160
['NADPH', 'TPNH', 'beta-NADPH'] --> NADPH

Figure 7: Sample results of Aron et al. (2007) (top) and Scott and Viola (1998) (bottom) (best viewed in color).

The difference, however, is that in these cases, previously *unknown* information is extracted, based on criteria that often take the form of templates in which potential slot fillers are defined in terms of semantic types (e.g. ENZYME) and (in the case of numerical values), ranges. In DB-to-document backprojection, in contrast, the expected information is explicitly known, fully specified, and 'only' needs to be located on the string level. Therefore, in contrast to a lot of the related work mentioned above, methods involving semantic similarity (like BioBERT (Lee et al., 2020)) are not necessarily superior to simple string matching when DB-to-document backprojection is concerned.

7 Conclusions & Future Work

In this paper, we introduced, defined, and performed some preliminary experiments with *DB-to-document backprojection*, a novel scientific document processing task. Our motivation for attempting this task comes from the requirements of a biocuration project, and from our idea to re-purpose previously unused (or rather *under*-used) data to advance biocuration methods. The focus of this initial paper was mostly on motivation, on a definition of the task and its functional parameters, and on the development of a better understanding of the effects and interactions of these parameters. For the latter, we performed some simple experiments and analysed the results. Rigorous evaluation, however, was not attempted, and, what is more, our results showed that defining what it means for a backprojection to be correct is difficult. While a quantitative evaluation remains difficult, a qualitative inspection of backprojection results clearly showed that the answer to our original research question is a positive one, which is the main result of this paper. Our findings regarding the role of color highlighting for backprojection, on the other hand, are somewhat mixed: While our system is able to detect highlighted tokens with high accuracy, appropriate ways to integrate this information into backprojection still need be be explored much further. The approach of *requiring* highlighting for matching, while not disproved yet, might be too strict, and alternative strategies will be evaluated in future work, for which our system and data sets provide a valuable basis. Additional future work includes the following: The optimization heuristics against OCR errors, although already shown to be effective in practice, are far from complete, and should be improved by handling additional cases of OCR errors, and other spelling variants. Also, as suggested by one reviewer, XML versions of papers from e.g. PubMed could be used to inform the backprojection task, which might also include automatic correction of OCR errors. Future work will also include the creation of an annotated dataset by inspecting the automatic results and storing correct highlighting in the extracted XML. Ideally, this should be done with the help and feedback of domain experts. Finally, the system will be applied to our full data set of $6,000+$ documents, which will yield a stronger data basis for analysis.

Apart from the obvious use cases like quality assurance in biocuration (and other fields where information is manually extracted from documents), and support for users of biological databases, both by means of visualizations, we also envisage several other potential applications for DB-to-document backprojection. These include creation of multimodal training data for page-topology-based document understanding systems like Katti et al. (2018), creation of input for empirical studies on document structure (distribution of information in scientific documents), and others.

Acknowledgements

This work was done as part of the project DeepCurate, which is funded by the German Federal Ministry of Education and Research (BMBF) (No. 031L0204) and the Klaus Tschira Foundation, Heidelberg, Germany. We thank the anonymous reviewers for their helpful suggestions.

References

Marwa Abdelhakim, Eunice McMurray, Ali Raza Syed, Senay Kafkas, Allan Anthony Kamau, Paul N. Schofield, and Robert Hoehndorf. 2020. Ddiem: drug database for inborn errors of metabolism. *Orphanet Journal of Rare Diseases*, 15(1):146.

Marco Anteghini, Jennifer D'Souza, Vítor A. P. Martins dos Santos, and Sören Auer. 2020. Representing semantified biological assays in the open research knowledge graph. *CoRR*, abs/2009.07642.

Zachary D. Aron, Pascal D. Fortin, Christopher T. Calderone, and Christopher T. Walsh. 2007. Fenf: Servicing the mycosubtilin synthetase assembly line in trans. *ChemBioChem*, 8(6):613–616.

George Buchanan and Fernando Loizides. 2007. Investigating document triage on paper and electronic media. In *Research and Advanced Technology for Dig-*

ital Libraries, pages 416–427, Berlin, Heidelberg. Springer Berlin Heidelberg.

Virginia Clinton. 2019. Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, 42(2):288–325.

Wei-Chih Huang, Hsin-Tzu Huang, Po-Yuan Chen, Wei-Chi Wang, Tai-Ming Ko, Sirjana Shrestha, Chi-Dung Yang, Chun-San Tai, Men-Yee Chiew, Yu-Pao Chou, Yu-Feng Hu, and Hsien-Da Huang. 2020. Svad: A genetic database curates non-ischemic sudden cardiac death-associated variants. *PLOS ONE*, 15(8):1–14.

International Society for Biocuration. 2018. Biocuration: Distilling data into knowledge. *PLOS Biology*, 16(4):1–8.

Peter Karp. 2016. Can we replace curation with information extraction software? *Database: The Journal of Biological Databases and Curation*, 2016.

Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469, Brussels, Belgium. Association for Computational Linguistics.

Maja Köpper, Susanne Mayr, and Axel Buchner. 2016. Reading from computer screen versus reading from paper: does it still make a difference? *Ergonomics*, 59(5):615–632. PMID: 26736059.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Robin Liechti, Nancy George, Sara El-Gebali, Lou Götz, Isaac Crespo, Ioannis Xenarios, and Thomas Lemberger. 2016. Sourcedata - a semantic platform for curating and searching figures. *Nature Methods*, 14:1021–1022.

Aldo Lipani, Florina Piroi, Linda Andersson, and Allan Hanbury. 2014. Extracting nanopublications from IR papers. In *IRFC*, volume 8849 of *Lecture Notes in Computer Science*, pages 53–62. Springer.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. DocVQA: A dataset for vqa on document images.

Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, Lara O'Donnell, Genie Leung, Rochelle McAdam, Frederick Zhang, Sonam Dolma, Andrew Willems, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. 2019. The BioGRID interaction database: 2019 update. *Nucleic Acids Research*, 47(D1):D529–D541.

Mary Ellen Scott and Ronald E. Viola. 1998. The use of fluoro- and deoxy-substrate analogs to examine binding specificity and catalysis in the enzymes of the sorbitol pathway. *Carbohydrate Research*, 313(3):247 – 253.

Sahar Vahdati, Said Fathalla, Sören Auer, Christoph Lange, and Maria-Esther Vidal. 2019. Semantic representation of scientific publications. In *TPDL*, volume 11799 of *Lecture Notes in Computer Science*, pages 375–379. Springer.

D Vishwanath, Rohit Rahul, Gunjan Sehgal, Swati, Arindam Chowdhury, Monika Sharma, Lovekesh Vig, Gautam M. Shroff, and Ashwin Srinivasan. 2018. Deep reader: Information extraction from document images via relation extraction and natural language. In *Computer Vision - ACCV 2018 Workshops - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers*, volume 11367 of *Lecture Notes in Computer Science*, pages 186–201. Springer.

Ulrike Wittig, Maja Rey, Andreas Weidemann, Renate Kania, and Wolfgang Müller. 2018. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Research*, 46(D1):D656–D660.

Ulrike Wittig, Maja Rey, Andreas Weidemann, and Wolfgang Müller. 2017. Data management and data enrichment for systems biology projects. *Journal of biotechnology.*, 261:229–237.

Wenyi Wu, Yan Wu, Dahui Hu, Yincong Zhou, Yanshi Hu, Yujie Chen, and Ming Chen. 2020. PncStress: a manually curated database of experimentally validated stress-responsive non-coding RNAs in plants. *Database*, 2020. Baaa001.