IUPAC pK_a Data Digitization Report

Conducted by:

Green Group @ MIT

Author	*Email*
Jonathan Zheng	jonzheng@mit.edu
Olivier Lafontant-Joseph	olivj23@mit.edu

May 20, 2024

Table of Contents

1 Introduction

1.1 Summary of report

This report details the digitization and curation of data presented in three reference books published by the International Union of Pure and Applied Chemistry (IUPAC):

1. **Serjeant**: Ionisation Constants of Organic Acids in Aqueous Solution; E P Serjeant and Boyd Dempsey; Oxford/Pergamon (1979) (Oxford IUPAC chemical data series)
2. **Perrin**: Dissociation Constants of Organic Bases in Aqueous Solution; DD Perrin; Butterworths (1965)
3. **Perrin Supplement**: Dissociation Constants of Organic Bases in Aqueous Solution, Supplement 1972; DD Perrin; Butterworths (1972)

We requested permission to scan and digitize theses references from the copyright holder IUPAC. IUPAC granted us the written permission on December 21, 2021, provided that the output data are reviewed by IUPAC and posted in an IUPAC GitHub repository and support the FAIR (Findable, Accessible, Interoperable, Reusable) data principles. With their permission, we digitized the reference books in-house. A digitized copy of Perrin (1965) was obtained from the Internet Archive using IUPAC's permission.

Additionally, we procured permission to digitize the information in the following three reference books:

1. **Kortum**: Dissociation Constants of Organic Acids in Aqueous Solution; G Kortum, W Vogel and K Andrussow; Butterworths (1961)
2. **Perrin Inorganic**: Dissociation Constants of Inorganic Acids and Bases in Aqueous Solution; D D Perrin; Butterworths (1969)
3. **Izutsu**: Acid-Base Dissociation Constants in Dipolar Aprotic Solvents; Izutsu, K; Blackwell (1990)

Data from these books are excluded from this report since the cleanup and curation are still in preliminary stages.

1.2 Other pK_a datasets

Other pK_a datasets exist and have been used in data science applications. However, they often include deficiencies in their reliability, accessibility, or quality. Below are listed several such datasets:

- **QSAR Toolbox** - large collection of aqueous data; dubious data sourcing, though some references are listed; no evaluation of uncertainty or rich metadata.
- **DataWarrior** - includes SMILES and presented in spreadsheet format, but dubious data sourcing, no evaluation of uncertainty; very little metadata.
- **OpenEye pK_a Prospector** - not open source.

- **iBonD pK_a** - very large and includes multiple solvents, but only presented in searchable format (e.g. difficult to access entire dataset), and lacking in contextual information, e.g. temperature, multiple dissociations.
- **OCHEM dataset** - crowdsourced; may not be accurate, or may be inconsistent e.g. different pK_a conventions; lacks rich information.

1.3 Definitions

1.3.1 Definition of terms used in this report

1. **pK_a**: Any type of dissociation constant: pK_A, pK_B, pK_{AH}
2. **T**: Temperature, most often in degrees C
3. **SMILES**: Simplified molecular-input line-entry system (SMILES string); one computer-friendly way of representing a chemical structure

1.3.2 Definition of terms used in the dataset

1. **18+1**: In the `original_T` column, the `+` is used to indicate "'plus-or-minus" ($\pm$).
2. $\sim$: approximately, i.e. `I~0.04` means I is approximately equal to 0.04
3. **10-4**: 10^{-4} (the exponentiation convention is to leave out the carat sign if there is a 10 next to a negative number)
4. **H0, Ho (and variants)**: Acidity function; other variants include `H_`, `HR`, etc.
5. **I**: ionic strength
6. **m**: concentration in mole per 1000g of water
7. **c (or C)**: concentration in mole L^{-1}
8. κ: specific conductance of water used for data entry, in 10^{-6} Ω^{-1} cm^{-1}
9. **(assumed)**: Some entries in the column "original_T" include this phrase. This refers to the case where the data entry does not explicitly label the temperature, but is implied to be carried over from a preceding entry. For transparency, these instances are labeled with this `(assumed)`.

2 Information about the dataset

The dataset includes:

1. **24,222** pK_a entries
2. **10,624** unique molecules (based on unique InChI strings)

The dataset is organized with **21** headers:

1. `unique_ID`: Unique identifier, combining the `source` and the `entry_#` to create a code that matches to each chemical species from each source
2. `SMILES`: SMILES string translated from the IUPAC names provided in the original reference work
3. `InChI`: InChI string derived from the `SMILES` strings, using RDKit API
4. `pka_type`: Type of dissociation constant. Examples: `pKaH1` = first conjugate acid's dissociation; `pKb` = basic dissociation constant. The vast majority of entries will be of the form `pKaH`, `pKa`, or `pKb`, but there are some exceptions in the reference works that include parentheses to identify an unusual protonation site or structure, e.g. `pK(indole-ring)` is pK for protonation on indole ring. (This convention may be changed in a later version.)

   Many amphoteric molecules species are also present in this dataset, for which several pK_a values are reported. The original reference works do not often distinguish which values are acidic and which are basic. This distinction can be automatically made if only two acidities are reported, in which case the lower value is assumed to be basic and the higher as acidic. For all entries with more than 2 pK_a values that are potentially amphoteric, we manually examined the chemical structures to determine the labels. Out of the original data corpus, **1,275** entries had ambiguous labels we could not manually assign, so we excluded them from the dataset. Future work may include resolving the labels for that missing set.
5. `pka_value`: the pK value
6. `T`: temperature (°C)
7. `remarks`: Comments for this specific datapoint.
8. `method`: Code for the method used to determine this datapoint.
9. `assessment`: IUPAC critical evaluation of the data uncertainty.
10. `ref`: Code for the reference to the original source of this datapoint
11. `ref_remarks`: Remarks pertaining specifically to the reference for this entry; e.g. a phrase like `Thermodynamic quantities are derived from the results`
12. `entry_remarks`: Remarks pertaining to all entries for this chemical species, usually pointers to other data sources (e.g. `Other values in B10`)

13. `original_IUPAC_names`: IUPAC identifier for the chemical species originally presented in the reference work (with a few exceptions, see the report for details)
14. `name_contributors`: The names of the methods that produced viable SMILES names from the IUPAC translation.
15. `num_name_contributors`: The number of methods that yielded a successful SMILES translation.
16. `original_IUPAC_nicknames`: Secondary names identified from the IUPAC identifier for the chemical species originally presented in the reference works
17. `source`: Name of the reference book: `perrin`, `perrin_supp`, or `serjeant`
18. `pressure`: Pressure value with units (a handful of entries include very high pressure entries which might yield unexpected results if not filtered, so this column is added to help filter these out)
19. `acidity_label`: Descriptor indicating whether the pK is an acidic (A), conjugate acid (AH), basic (B), or "other" dissociation constant
20. `original_T`: Displays the original temperature if it was corrected for purposes of standardization. (In the "T" column, room temperature was converted to 25 °C, and any "approximate" temperatures such as ∼ 20 were reported without the ∼ sign)
21. `solvent`: Solvent information, if parsable from `remarks`

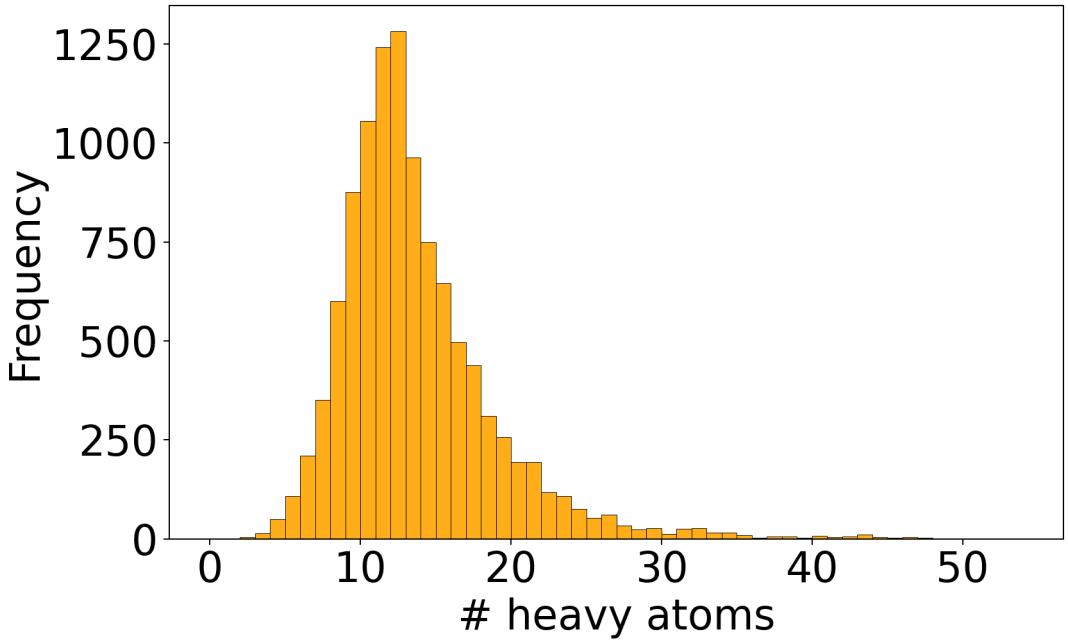

Figure 1: Histogram showing the distribution of atoms in the dataset based on number of heavy atoms. The distribution is right-skewed with most species falling between 8-18 heavy atoms.

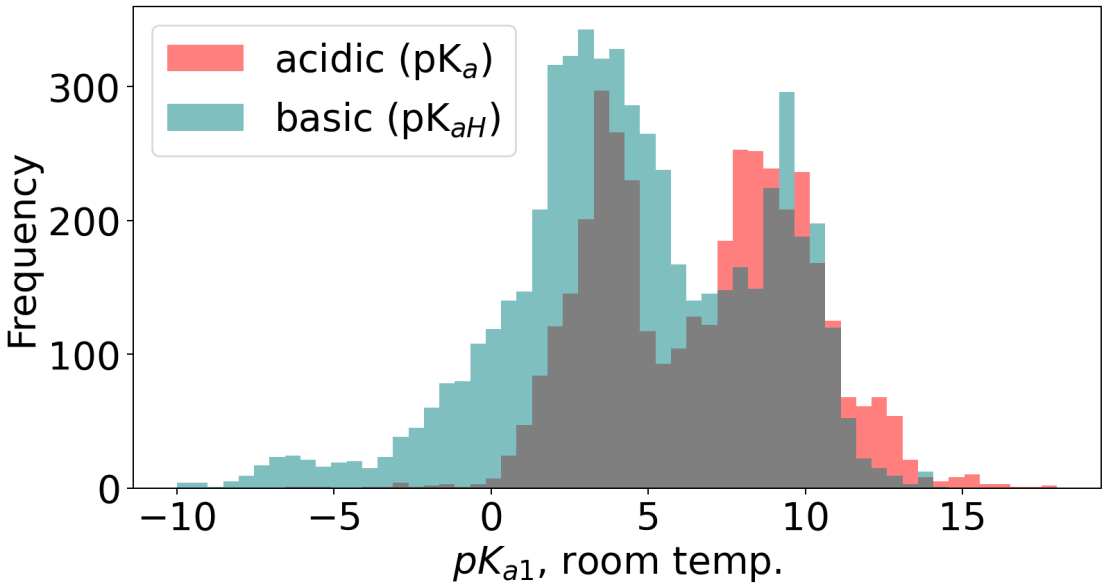

Figure 2: Histograms showing the distribution of near room-temperature first dissociation pK$_a$ values in the dataset. Both sets of acidic and basic pK$_a$ values peak near a value of 4 and show a secondary peak near 10. pK$_a$ values plotted here are averages for all values found near a temperature (i.e. molecules are not repeated per acidity type).

The peak locations in the pK$_a$ histograms match those reported by other pK$_a$ datasets.

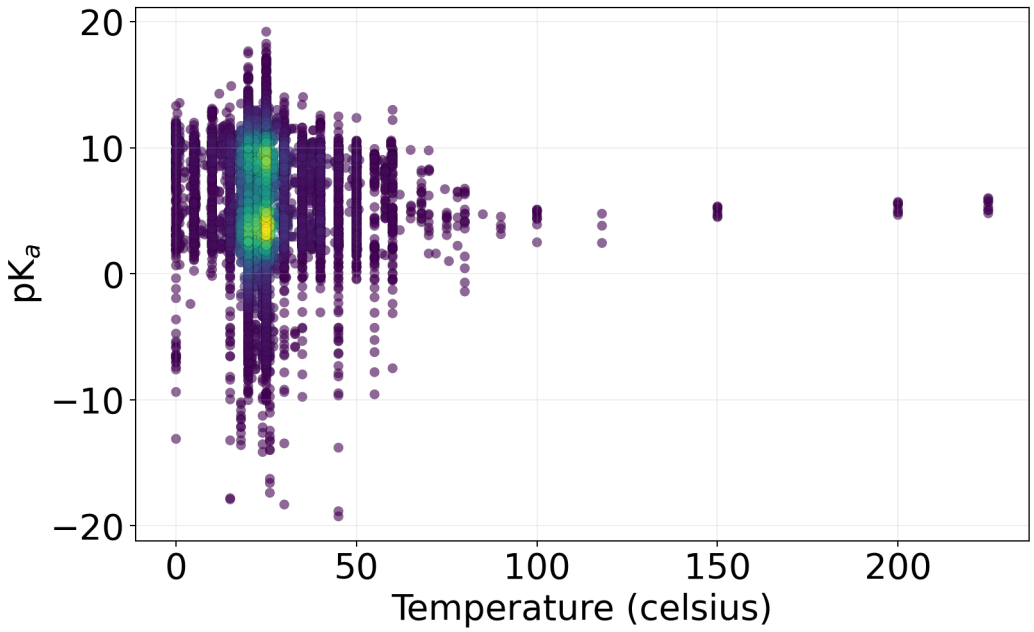

Figure 3: Plot showing the pK$_a$ and temperature range across the dataset. Pressure is not controlled for or represented in this diagram. The majority of temperatures are at or near room temperature.

3 Methods

The general workflow for creating the dataset was as follows:

1. Scan the reference books.
2. Use **optical character recognition (OCR)** to convert the scanned images into text loosely ordered into tables.
3. Organize the loosely-ordered tables into highly structured human-readable tables.
4. Check the tabulated data for errors manually.
5. Add machine readable identifiers (SMILES) to each entry and convert the human-readable tables to machine-readable tables.
6. Write scripts to identify suspicious entries that need verification or correction.

This report documents the details of these steps.

3.1 Making scans of the reference books

With IUPAC's permission, the reference books were scanned by the author in the MIT Libraries, or a digital copy was obtain from the Internet Archive (archive.org).

3.2 Converting scanned images into text using OCR

The scanned images were then uploaded to AWS S3 cloud storage. The three reference books were OCR'd using Amazon Textract.

3.2.1 Benchmarking OCR

To maximize accuracy, we attempted to improve the image quality before performing OCR. We found that the quality was improved by de-skewing the scans (such as through using software packages that implement `unpaper`) and removing image artifacts. Both of these tasks were completed using open-source code `ocrmypdf`.

We also benchmarked the color and cropping of the image. We compared the quality of the OCR to the case where the image was grayscaled, and found that in some cases the quality improved whereas in others the quality decreased. Observing no consistent pattern in improvement, we decided to stick with the full-color OCR. We also checked whether cropping the scans to ignore the book margins would improve OCR quality, but also observed no consistent improvement.

The rationale behind these changes is that the OCR software we used is trained on a huge corpus of scanned images, almost all of which are not books with tabulated data: for instance, receipts, signed documents, paperwork, etc. So, we had hoped that modifying the scans to resemble the training scans would improve quality. Though we did not see a consistent improvement in quality, we did see occasional improvements, and believe that this other

projects that attempt this sort of workflow may benefit from doing a similar benchmarking study.

We also attempted to benchmark OCR software. We found that freeware (i.e. Tesseract implemented in `ocrmypdf`) generally performed worse than Amazon Textract. We did not investigate other private OCR software such as Google Cloud Vision or Microsoft Azure OCR.

3.3 Organizing scanned text into human-readable tables

Once the scanned images were converted into tables and then loosely organized into tables, we further processed those loosely-organized tables into a highly-structured format. This format reflects the way that information is conveyed in the original IUPAC texts and meant to be read by a human. At this stage, the column headers include:

1. Entry number, e.g. `2001`
2. IUPAC name(s)
3. pK_a type e.g. `pKa1`
4. pK_a value e.g. `7.3`
5. Remarks
6. Method
7. Reference
8. Data evaluation
9. Reference remarks (Information presented in the text that applies to all data for this species falling under just this reference)
10. Entry remarks (Information presented in the text that applies to all data for this species, across all references)

We first performed a preliminary manual organization step by ensuring that all entry numbers lined up in the first column. We developed an algorithm that sorts through the loosely-organized tables and formats them. The details of the algorithm are presented in the Appendix.

3.3.1 Special considerations

When using spreadsheet editors, particularly LibreOffice Calc, the default way that text is read in might result in data corruption. For instance, a number like `16.0` might be truncated to `16` if the formatting type is Number. Hence, all data should be formatted as "Text", rather than "Standard".

3.4 Checking tabulated data

Each of the organized tables were split into subsections corresponding to roughly 50 pages of the original reference book. Various members of the Green Group were trained to double-check the image scans with the tables, and then checked many of the values in the Serjeant book. The author of this document then triple-checked that all errors were corrected. The author was the sole reviewer of the Perrin and Perrin Supplement books.

In each checking process, the scanned images are compared to the tables (since their layout is nearly identical). Each table was edited to be read as similarly as possible to the reference work.

3.5 Organizing human-readable tables into computer-readable tables

Once all pages were double-checked at least once, the pages were concatenated into yet another table format. The cleaned data was read by another script that extracts all synonym names provided by the verbatim textbook name. For instance, the provided name `Benzaldehyde, 2-hydroxy- (Salicylaldehyde)` is represented in the dataset as the two names `Benzaldehyde, 2-hydroxy-` and `Salicaldehyde`. These names are all fed into OPSIN, the Chemical Identifier Resolver, Pubchem, and Chemaxon's Molconvert software. Taking all such names into consideration, if these software differed in their translation, the SMILES were labeled as **Inconsistent**. If no software was able to translate the name, then the SMILES was labeled as **Missing**. Otherwise, the SMILES was considered **Converged**. If one SMILES string indicated stereochemistry while the others did not, then the string with stereochemistry was used as the converged SMILES. All missing and inconsistent SMILES strings were noted and separated from the converged SMILES set.

The dataset was manually screened several times to correct any remaining irregularities found in the dataset. Extra care was taken to detect and fix abnormal entries, such as entries with abnormally high or low pK_a values, missing locants in the names, common typos in the `remarks`, and so on. The data was also organized into the headers that are presented in this version of the dataset. For instance, the `remarks` were scanned for key words that indicated the pressure of the entry, used to populate the `pressure` header. The header `acidity_label` was determined by comparing the pK_a source (Perrin and its supplement are always basic or conjugate acids) and the reported `pka_type`.

3.6 Final checks on dataset

We checked to make sure that all entries were included by checking to make sure every number between the first and last entry numbers was recorded in the database. We noted that some of the reference works did not meet this criterion in the original print:

1. Serjeant: `[2306, 5189, 2310, 4380, 4381]`

We employed a variety of programmatic checks and standardizations:

1. Flag entries that had missing pK_as and missing references.
2. Calculate the pK range and distribution, and check any outliers.
3. Label potentially ambiguous pK_a types with "?" in the `pka_type` and `acidity_label` (arising from potentially amphoteric molecules). If multiple acidities are reported, then the molecule is potentially polyprotic (in which case the original acidity labels are correct) or amphoteric (in which case they are incorrect). In Perrin, amphoteric molecules are distinguished from polyprotic molecules, but in Serjeant they are not. Therefore, for any molecule that *could* be amphoteric based solely on data presentation rather than molecular structure, we labeled the pK_a type with a question mark. In the current form of this dataset, such entries containing question marks were removed.
4. Standardize all `assessment`, `pka_type`, and `remarks` to contain as similar language as possible (e.g. standardizing "Very Uncert." and "V. Uncert.")
5. Flag suspicious names for `pka_type`.
6. Flag entries that do not have a `reference`.
7. Search for salts in SMILES. We found that several salts were present in the dataset, but elected to keep them in the final set.
8. For molecules with only two listed pK_a value that are known to be amphoteric, standardized the lower value to be the basic dissociation and the higher one to be acidic. Manually checked all other amphoteric molecules to ensure correctness of the pK_a type.
9. Standardize temperature values so that they can be parsed as numbers. (Note: Did not standardize temperature values that begin with "<").
10. Modify the SMILES strings to be stereospecific when appropriate (i.e. based on the name or `pka_type`)

The full dataset, including species with missing and inconsistent SMILES, totaled 27,026 entries. Next, the dataset was pruned to exclude those missing SMILES strings and entries that only had one programmatic check without manual verification, resulting in a subset totaling 25,498 entries. Finally, the dataset was pruned again to include only entries with unambiguous pK_a types (i.e. pK_a vs. pK_{aH}), resulting in the dataset of 24,222 entries. This means approximately 3,000 entries could not be digitized with sufficient confidence for this work.

Ongoing work is in translating the remaining IUPAC names to SMILES. Next steps also include developing programmatic checks on the dataset to further detect outliers (for example in pK_a values depending on functional groups present).

4 Acknowledgements

Special thanks to:

- Ye Li, who has consistently supported this project, provided feedback to this report, and brought together the many people who have influenced this work;
- Leah McEwen, for key insights during the digitization and parsing process, coordinating countless meetings, and bringing together so many IUPAC volunteers who were instrumental to this work: Stuart Chalk, David Shaw, Glenn Hefter, and executive director Dr. Lynn Soby;
- my doctoral advisor, Professor Bill Green, and members of the Green group at MIT for their help with checking entries, including: Yunsie Chung, Xiaorui Dong, Michael Forsuelo, Kevin Greenman, Esther Heid, and Hao-Wei Pang;
- PubChem scientists including Evan Bolton and Jeff Zhang for their support in translating the reference names to SMILES strings;
- Prof. Ivo Leito and the group at the University of Tartu;
- IUPAC, for granting permissions to digitize the reference books and providing guidance with curating the dataset;
- Internet Archive for providing the digitized copy of Perrin (1965) with IUPAC's permission.

This project also utilized proprietary software: Amazon Textract to perform the OCR, and ChemAxon Molconverter (`https://chemaxon.com`) to aid in the translation of IUPAC names to SMILES strings.

We also acknowledge use of open-source packages: `py2opsin`, `unpaper`, `OCRmyPDF`, `camelot`, `tabula`.

Appendices

A Common errors

A.1 Errors/typos in OCR

Systematic errors with the OCR quality were observed:

1. Molecular formulas such as `C3H7O` were often completely wrong.
2. `l` replaced by `1`
3. `1` replaced by `l`
4. `O` replaced by `0`
5. Missing `1` in chemicals such as `1,3-dichloro...`
6. Periods parsed incorrectly as commas.
7. Spaces erroneously inserted throughout text, particularly near decimals.
8. `pK1` parsed incorrectly as `pk,` or `pky` or `pkg`
9. `4` parsed as `s`
10. Tables inconsistent; data were sometimes shifted from one column to another
11. Rows near the end of the pages were missing (particularly if the page is warped or cropped near the bottom).
12. Entire pages were sometimes missing; in this case, they were manually transcribed.

A.2 Deviations in digital dataset from original typesetting

There were a number of cases when the formatting in the book deviates slightly from the formatting chosen in this dataset.

Serjeant:

1. Entry 2076 uses the terms `pka`, `pkb`, `pkc`, and `pkd`. These were changed to `pk1`, `pk2`, `pk3`, and `pk4` to avoid confusion with the base dissociation constant, pK_B.
2. Various entries have "subtables", or tables embedded within the reference text that contain data information varying across a variable, usually temperature but also sometimes ionic strength and pressure. These are parsed as separate entries in the dataset if they vary along temperature or pressure. Otherwise, for entries like Entry 2835, if the information varies along ionic strength, then the info is kept as a `Entry Remark`.
3. Entry 5283 has a superscript that is represented with a carat e.g. `P^1,P^2` to represent P^1,P^2.

Perrin:

1. Entry 121 includes odd nomenclature, which was parsed as `From ref and extrapolation to m=0`
2. Some entries include equations for a certain temperature range. For instance, `pK=2.524 + 1563/T (T in K); t=10 to 50 deg.` in entry 121. Other equation examples: entry 399, 946, etc. These are reported as individual pK_a entries. Other examples include Serjeant entries 7622, 7629.
3. entries 252 and 253: the entries appear with `Meso` and `Racemic` in the name. We elected to move those into the `Remarks` section for each entry.
4. Entry 382: The book name is `Cycloheptane, (trans?)-1,2-diamino-`, but we instead name it as `Cycloheptane, 1-2-diamino-` and make a note for the Remarks that the entry possibly refers to the trans form.
5. In some cases, the textbook uses square brackets when rounded parentheses would do. In this work, when square brackets are used *just* to cluster groups (and not to identify locants, for instance on a bicyclic molecule) the two formats are used interchangably; for instance, Entry 485 (and similar): `Aniline, 4-[(2-diethylamino-4-methyl)pentyloxycarbonyl)]-`.
6. See entries 610 and 613: Reactions are denoted as `HL=H+ + L-` to refer to the dissociation of HL into a proton and the L^{-1} anion.
7. See entries 924-929: These are not species names, but refer to references where further information can be found. Also see entry 1168, 1417, 1531, 2123, 2222, 2232, 2851, 2853-2855, 3767 (not an exhaustive list).
8. Some entries have chemical diagrams in the textbook that can be used to infer the SMILES, e.g. entry 1254, where OPSIN fails. (triazole entries around 1848; 2771-2773, 2777; 2837-2846; 3699-3719; 3721-3741). These entries were not successfully translated from IUPAC to SMILES.
9. Entries 142, 1433, and 1464 are stylized as Δ^2 but in the dataset are represented as $\Delta 2$
10. Entry 2157 and others include substrings `5-(or 6)-` to refer to the fact that the species tautomerizes so that a proton at either the 5 or 6 site are identical. These substrings were respresented without the "or" substring, e.g. `5(6)-` instead of `5-(or 6)-`. In these cases, a transcription note was also included e.g. 'Note from transcription: Naming for `Benzimidazole, 2-methyl-5-(or 6)-nitro-`. The SMILES strings for these entries may encounter difficulties with OPSIN, e.g. entries 2157, 2160, but can be easily manually found. A similar issue was encountered for entries 2737-2742; `1(3)H-Naphth...` was replaced with `1H-Naphth...` and a note was provided.
11. Entry 2483a: this new entry was added based on the pK_a of one of the pK_as listed for Entry 2483 (a special case; but the Remarks originally indicated the pK_a was for a different species)

12. Entries 2599-2646 use an outdated nomenclature, `pyrazolo(5',4'-4,5)pyrimidine`. Based on the structures provided in the textbook, these were replaced with `Pyrazolo[3,4-d]pyrimidine`. According to this convention, the following loci are mapped (left numbers is original nomenclature, right number is mapping with updated nomenclature):
    $1' \rightarrow 1$
    $2' \rightarrow 2$
    $3' \rightarrow 3$
    $6 \rightarrow 4$
    $1 \rightarrow 5$
    $2 \rightarrow 6$
    $3 \rightarrow 7$
13. Entries 2594-2598 refer to a slightly different species, "pyrazolo(4',5'-4,5)pyrimidine" which I replaced with "Pyrazolo[4,3-d]pyrimidine". As before, the mapping is:
    $1' \rightarrow 1$
    $2' \rightarrow 2$
    $3' \rightarrow 3$
    $3 \rightarrow 4$
    $2 \rightarrow 5$
    $1 \rightarrow 6$
    $6 \rightarrow 7$
14. Entries 2653 and 2654 refer to a slightly different species, `1,2,3-Triazolo(5',4'-4,5)pyrimidine`. Entry 2653's entry was replaced with `8-Purine` based on the provided textbook image, but declined translating entry 2654 as of the first version of this dataset.
15. Entry 3138 includes the `[Oxidized]` substring; this was removed from the name and inserted into the Remarks.
16. Entry 3215 cites `pK` as the pK type, in contrast to how pK values are reported for other entries in the textbook (usually either as `pK1` or with no listed pK type). This was assumed to be the first dissociation `pk1`.
17. Entries 3256-3260 include the `(DL)` prefix; the parentheses were removed e.g. `DL-Phenylalanine amidoxime`.

Perrin Supplement:

1. Entry 4253a "tricyclo[3.3.1.1[3,7]]decane, 1-amino-" has a superscript in it; the superscript numbers are denoted in the inner pair of square brackets, "[3,7]".
2. Entries 5563-5571 feature a Δ^2 nomenclature which is indicipherable using OPSIN. Entry 5563 includes an example molecule diagram, which might be usable to figure out how to parse the rest of the SMILES. Same for entries 6490-6496.
3. Entries 6644a-6644c originally appear as contextual information in Entry 6644's `Remarks`, and are separately meted out into individual entries here.

Some inconsistencies may exist with the exponentiation formatting. In most cases, the formatting leaves out the carat sign, e.g. `10-4` refers to 10^4. Similarly, most entries use `Ho` rather than `H0` or `H_0` to refer to H_o (Ho is generally preferred, but some entries such as 634 use H0). These may become standardized in a future version of this dataset.

A.3 Errors in name translation to SMILES

We noticed that chemical names with `'` tended to fail in the OPSIN name translation steps. Further investigation into common failure modes is currently underway.

A.4 Possible errata in the original reference texts

Perrin:

Entries 740 and 741 have the same name but a different nickname (and are isomers). The same was observed for Perrin Supplement entries 4806 and 4807.

Some species had question marks in their name: Perrin entries 2015, 2020, 2021. See also Serjeant 3012-3015.

Perrin Supplement:

Entries 4577, 5014 are missing a `Ref.`

Entry 5682 has a `5,5-tetramethylene-` fragment, but there should be 4 locants. Thus the SMILES string is indicipherable.

Entry 6358: One entry (pK=-7.03) doesn't have a proper pK label and has an unusually low pK value. Thus the pK type was transcribed as `pK(?)`.

Entry 6540: There are more left braces than right braces in
`1,3,5-Triazine, 2,4-diamino-6-[6'-(2',4'-diamino-1,3',5'-triazinyl]amino-`.

B Data formatting algorithm from text to tables

1. First, the data is preprocessed. Extraneous header rows, which repeat across each page, are detected by the presence of key words (`No.`, `Molecular formula`, `Remarks`, `Method`, and so on) and then removed. Entry numbers that leak into the molecular name are split into two cells (e.g. `2002Propanoic` into `2002`, `Propanoic`). Molecular formulas are also detected (by assigning each word a score based on how many element-based characters like C, H, O, and N are present and lowered based on how many word-like characters like commas and hyphens are present) and removed. This was done because the OCR for molecular formulas was very inaccurate so molecular formulas were simply removed.

2. Next, the data was organized into a custom container for each molecule. Each container contained sub-containers which contained the information for each line entry under each molecule. The input dataset is read line-by-line; if an entry number is detected in the first column, then a new container is created (representing a new molecule). Each subsequent row of text is stored in the container until a new entry number is detected.

3. For each line of data stored in the container, the script attempts to extract information based on the location of the information and the data type, and assign them to sub-containers. To do this, the the containers are iterated through several times. Containers initially store the disordered raw text as lists; the first list is the first row of raw data, the second is the second row of raw data, and so on. Each row is iterated in the following fashion. In the first subroutine, the script searches for IUPAC names from the raw data.

   In the second iteration, the uncertainty, references, pK_a types, methods, and references are inferred from the raw data by matching to regular expressions (regex) or to a pre-defined list of possible values. The temperature is also inferred by searching for numbers that are either integers, large numbers, or decimals known to often occur as temperatures.

   In the third iteration, the script reads each line up to the lower of the point where the temperature was read or the fourth index in the row. The logic here is that the pK_a data is presented in columns preceding the temperature column, so pK_a data should be read first. If a decimal number is found, then it is stored as the pK_a value; simultaneously, if text in the format of `pK` is read, then that is inferred as the pK_a type.

   Next, the IUPAC name is inferred from whatever text is remaining in the first row.

   Any information that hasn't been parsed yet is concatenated and parsed in Remarks.

4. In a 'post-post-processing' step, the containers are again iterated through. In the first iteration, the algorithm infers whether a "subtable" is present by the presence of many pK_a values and presence of keywords such as "Variation with temperature". These are tables that occur within each reference work that contain several pK values across many operating conditions, often reported on one single line.

   Next, the script tries to clean up the IUPAC names. Sometimes information that belongs in the `Remarks` section finds its way to the `Names` section; the script detects these and moves them to the right section.

   Then the script checks if there are not more pK_a types than pK_a values. The program then checks to see if temperatures were mistakenly placed in another section. For instance, if the remarks contains no text and only a number, or if no pK_a value is recorded, or there are multiple recorded temperatures, then the script will fix these.

   Then, as one more check, if there is a row that only includes information on `Remarks` without any other information, then the `Remarks` info is placed into the next line.

Empty rows are detected and removed.

An algorithm also duplicates contextual information (temperature, reference) from one entry to the next one, if the former row has the rich info but the latter row does not. This is an assumption that all entries falling under the same entry retain the same information if that information is omitted from one line to the next. We believe this is the intended way that the reference books are meant to be read, but since there is an underlying assumption made here, we explicitly label the temperatures with "(assumed)" wherever possible.

5. The containers and their subcontainers are then parsed into rows in a .csv file.