HAL open science

Curated dataset of association constants between a cyclodextrin and a guest for machine learning

Gökhan Tahil, Fabien Delorme, Daniel Le Berre, Éric Monflier, Adlane Sayede, Sébastien Tilloy

To cite this version:

Gökhan Tahil, Fabien Delorme, Daniel Le Berre, Éric Monflier, Adlane Sayede, et al.. Curated dataset of association constants between a cyclodextrin and a guest for machine learning. Chemical Data Collections, 2023, 45, pp.101022. 10.1016/j.cdc.2023.101022 . hal-04056166

HAL Id: hal-04056166

https://univ-artois.hal.science/hal-04056166v1

Submitted on 15 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Curated Dataset of Association Constants Between a Cyclodextrin and a Guest for Machine Learning

Gökhan Tahıl[a,b]*, Fabien Delorme[a], Daniel Le Berre[a], Éric Monflier[b], Adlane Sayede[b], Sébastien Tilloy[b*]

[a] Univ. Artois, CNRS, Centre de Recherche en Informatique de Lens (CRIL), F-62300 Lens, France.

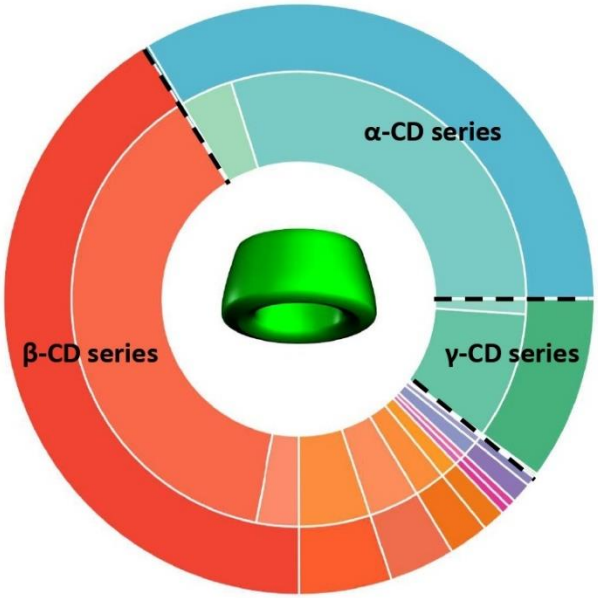
[b] Univ. Artois, CNRS, Centrale Lille, Univ. Lille, UMR 8181, Unité de Catalyse et Chimie du Solide (UCCS), rue Jean Souvraz, SP 18, F-62307 Lens Cedex, France.

gokhan.tahil@univ-artois.fr; sebastien.tilloy@univ-artois.fr

Abstract

Determining the association constant between a cyclodextrin and a guest molecule is an important task for various applications in various industrial and academic fields. However, such a task is time consuming, tedious and requires samples of both molecules. A significant number of association constants and relevant data is available from the literature. The availability of data makes the use of machine learning techniques to predict association constants possible. However, such data is mainly available from tables in articles or appendices. It is necessary to make them available in a computer friendly format and to curate them. Furthermore, the raw data need to be enriched with physicochemical information about each molecule and when such information does not allow to discriminate molecules, some additional data is needed. We present a dataset built from data gathered from the literature. The dataset contains both the original raw data from the articles and the enriched ones. We also provide the scripts used to curate and enrich the raw data.

Graphical abstract



CDs (native and modified) dataset
for association constants prediction

Keywords: Machine Learning, Association constant, Cyclodextrin

Specifications Table

Subject area	Chemoinformatics, organic chemistry, supramolecular chemistry
Compounds	Cyclodextrins (15 molecules) and selected guests (1767 molecules)
Data category	Association constants or Gibbs free energy for 1/1 cyclodextrin/guest inclusion complex
Data acquisition format	Extraction from publications and online databases
Data type	Raw and Structured Enriched Data
Procedure	Experimental association constants were extracted from various publications and values with large uncertainties were removed
Data accessibility	Data files URL: https://doi.org/10.5281/zenodo.7575539 Script files URL: https://doi.org/10.5281/zenodo.7575579

1. Rationale

Native cyclodextrins (CDs) are cyclic oligosaccharides consisting of six (α-CD), seven (β-CD), or eight (γ-CD) α-D-glucopyranose units and are industrially synthesized by enzymatic conversion of starch. CDs possess an outer hydrophilic surface and inner hydrophobic cavity. In this cavity, a guest can be included to form an inclusion complex. To vary their size and shape, native CDs can be modified by etherifying some or all hydroxyl groups by various substituents (R) (Figure 1).

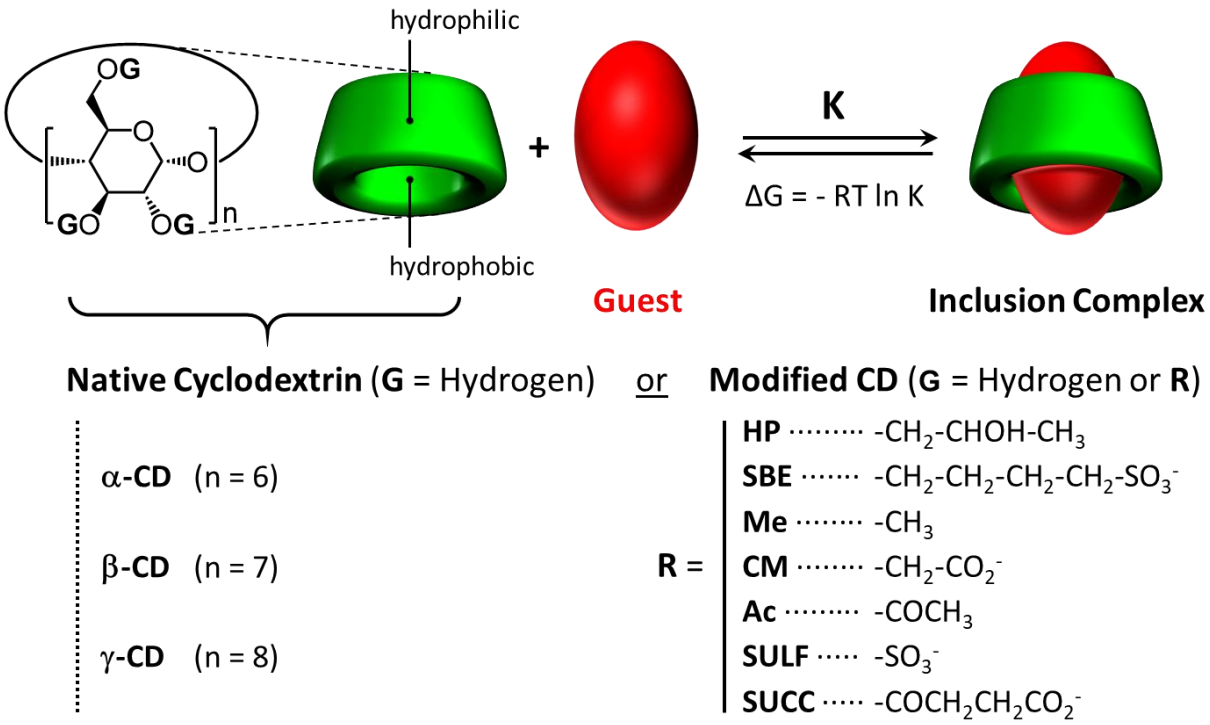


Fig. 1. Representation of an inclusion complex formation between a CD and a guest.

These are referred to modified CDs. Due to their ability to form inclusion complexes, CDs have applications in various industrial and academical fields [1,2]. The affinity between a CD and a guest is quantified by the association constant (K). Its determination is generally performed using analytical techniques such as NMR [3], UV spectroscopy [4] , isothermal titration calorimetry [5] or others, which are time consuming, sometimes tedious and often requires samples of both molecules. In this context, the prediction of the association constant from physicochemical features is appealing. An interesting alternative is to use machine learning approaches. Indeed, the application of machine learning to chemistry has increased

significantly in recent years [6,7]. These methods are also becoming very popular for the determination of the CD/guest association constant [8–11]. Their first requirement is the availability of a quality dataset. However, in some publications, the dataset used is not publicly available, the experimental conditions are not presented or the data format not easily reusable for other machine learning studies. For example, (i) 233 data are available but only for β-CD or sulfobutylether-β-CD [8] (ii) 1320 data are used including 8 different CDs but the dataset is not available [9], (iii) 1654 data are available included 16 different CDs acquired from the Cyclodextrin knowledge base (8534 collected data) unfortunately no longer available [12] and (iv) 280 data are available from the BindingDB community [11,13].

In this data paper, we propose a curated dataset of association constants. The data come from different articles/sources from the years 1963 to 2021 [10,11,14–19]. For each CD/guest pair (stoichiometry 1/1), the values of association constant K (or Gibbs free energy ΔG with $\Delta G = -RT\ln(K)$), pH, temperature (T), and the source of data are presented in the data files.

Such data can hardly be used as is because machine learning models are often built using numerical data, which should discriminate the CD/guest pairs. Non numerical data such as a molecule name must be replaced by dedicated physicochemical properties. When such data is not sufficient to discriminate the molecules, additional properties must be provided. In our context, we provide abstract additional features inspired by the state of the art in text processing, i.e., Word2Vec [20]. We also made sure that molecules are always represented the same way independently of their representation in the original source. However, using structured data is not enough to produce a high-quality predictor. Indeed, some values from the datasets may be erroneous for various reasons (experimentation, reporting, data collection, etc.). We made our best to detect and fix such erroneous values.

2. Procedure

2.1 Software environment

The dataset was built using scripts in python version 3.8.8. The raw data contained in PDF articles or data appendices were collected thanks to the Camelot library[1] version 0.8. In some sources, the data were provided as Structure Data Format[2] files. The RDKit library[3] version 2021.03.5 could read that format. The data in the Binding database were shared on web pages and the Requests library[4] version 2.26.0 was used to extract them automatically. The cheminformatics libraries PubChemPy[5] version 1.0.4 and RDKit were used to retrieve or compute physicochemical properties. The Pandas library[6] version 1.3.4 and NumPy[7] version 1.20.3 were used for tabular data manipulation, to save the data in a structured form and to ensure reproducibility in the development of the machine learning algorithms.

2.2 Dataset preparation

The basic data retrieved from the literature are the CD name, the guest molecule name, the value of association constant (or the Gibbs free energy) and when provided its error margin. Additionally, one can find the experimental conditions (generally pH and temperature). When pH and/or temperature were not reported, their values were set to pH=7 and/or T=25 °C. The final dataset contains 1767 guest molecules and 15 CDs (3 natives and 12 modified). The proportions of each CD are shown in figure 2A. The nature and the distribution of each CD are gathered in the Table 1. The values of Gibbs free energy for CD/guest inclusion complex were comprised in the range of -0.65 to -30.70 kJ/mol with a mean of -13.37 kJ/mol, as shown in figure 2B. The stoichiometry of CD/guest inclusion complex is 1/1. The guest molecules contain a large amount of structurally diverse organic compounds.

---

1 https://camelot-py.readthedocs.io/
2 http://biotech.fyicenter.com/resource/sdf_format.html
3 https://www.rdkit.org/
4 https://requests.readthedocs.io/
5 https://pubchempy.readthedocs.io
6 https://pandas.pydata.org/
7 https://numpy.org/

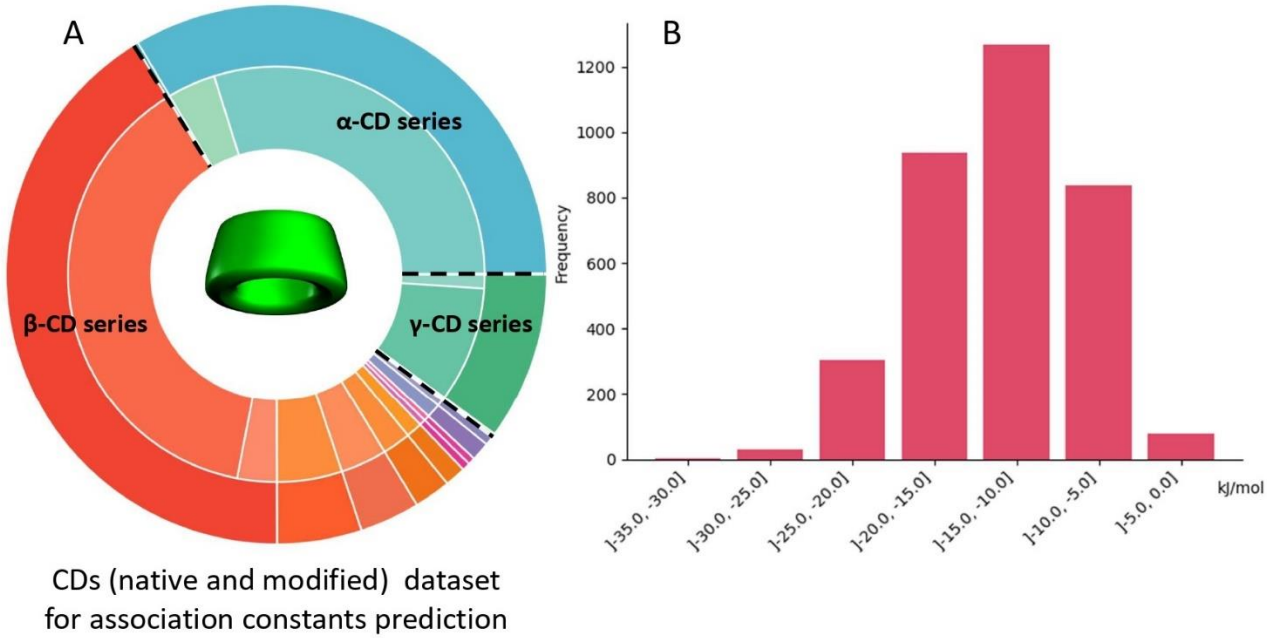


Fig. 2. (A) The distribution of data for each CD, the pie chart with a large radius shows the total data distribution in the raw data file. In the pie chart with a small radius, the light-colored parts show the eliminated data, and the dark parts show the data distribution in the final file. (B) The distribution of the association constant between CDs and guest molecules.

Table 1

Distribution of the various CDs in the dataset

α-CD series		β-CD series		γ-CD series	
Name	Number of Data	Name	Number of Data	Name	Number of Data
α-CD	1117	β-CD	1434	γ-CD	338
trimethyl-α-CD	8	hydroxypropyl-β-CD	190	hydroxypropyl-γ-CD	5
hydroxypropyl-α-CD	3	sulfobutylether-β-CD	135		
		methyl β-CD	82		
		dimethyl-β-CD	47		
		carboxymethyl-β-CD	42		
		acetyl-β-CD	19		
		trimethyl-β-CD	18		
		β-CD sulfate	16		
		succinyl-β-CD	5		

The main issue is to correctly identify the molecules. They are identified in articles using a name from a given taxonomy. Sometimes, the name itself was not sufficient to identify the molecule. In a publication [15], not only the name (as neutral form) but also its charge values are needed to identify the guest molecule. For example, 4-aminobenzoic acid data is shared in more than one data with charge value as 0, -1 and +1. With the nomenclature of the International Union of Pure and Applied Chemistry (IUPAC), these three molecules should be named differently from each other. For this reason, a manual inspection step was required when the charge associated with the molecule name did not match the charge available in the article. For instance, the IUPAC name of 4-aminobenzoic acid with charges 0, -1 and +1 are respectively 4-aminobenzoic acid, 4-aminobenzoate and (4-carboxyphenyl)azanium.

2.3 Molecular descriptors

The raw data have then been enriched with physicochemical properties. In earlier stages, all physicochemical properties were retrieved from the PubChem database [21]. It has subsequently been found that after PubChem's weekly updates, the properties retrieved for some molecules could be updated. Therefore, we limited the usage of PubChem for the retrieval of two information only: the PubChem CID (Compound ID) and the IsomericSMILES [22] value. The first information is the unique identifier for a molecule in PubChem database. It is to the best of our knowledge the most practical way to uniquely identify a molecule. The SMILES (Simplified Molecular Input Line Entry System) notation of a molecule is used to describe the structure and chemical bonds of that molecule. IsomericSMILES contains the isotopism and stereochemistry information of the molecule. However, those SMILES are not canonical, i.e., several IsomericSMILES can represent the same molecule. As such, we use the RDKit library to normalize the IsomericSMILES, i.e., to always represent a given molecule with the same IsomericSMILES.

For each host-guest pair, 7 features for the host molecule and 9 features for the guest molecule are computed by RDKit. Among them, 7 features are common: Topological Polar Surface Area (TPSA), Molecular Weight (MW), Complexity, Charge, HBondDonorCount (HBDC), HBondAcceptorCount (HPAC), HeavyAtomCount (HAC). **TPSA** is an estimate of the area (in $Å^2$) which is polar [23]. **MW** is the sum of all atomic weights of the constituent atoms in a compound (in g/mol). The **complexity** rating of a compound is a rough estimate of how complicated the structure is, seen from the point of view of both the elements contained and the displayed structural features including symmetry [24,25]. **Charge** represents the total charge of the molecule. **HBDC** and **HPAC** denote respectively the count of hydrogen bond donors and the

count of hydrogen bond acceptors. **HAC** is the number of atoms (heavy atoms) other than hydrogen atoms. The two following features are used only for guest molecules. The partition coefficient (logP) of a compound defines the ratio of its solubility in two immiscible solvents (octanol : water) and **MolLogP** is estimated by a property-based method [26]. The aromatic proportion (**AP**) is calculated by dividing the number of all aromatic atoms in a molecule by the number of all heavy atoms. For example, Table 2 shows the features generated by RDKit for some molecules.

Table 2

The features generated by RDKit for 7 different molecules

Guest	IsomericSMILES	TPSA (Å^2)	MW (g/mol)	Complexity	Charge	HBDC	HBAC	HAC	MolLogP	AP
acetonitrile	CC#N	23	41	24	0	0	1	3	0.53	0
hexylamine[1]	CCCCCC[NH3+]	27	102	23	1	1	0	7	0.81	0
isobutyric acid	CC(C)C(=O)O	37	88	56	0	1	1	6	0.73	0
(1R,2S)-ephedrine[1]	C[NH2+][C@@H](C)[C@H](O)C1=CC=CC=C1	36	166	222	1	2	1	12	0.30	0.5
(1S,2R)-ephedrine[1]	C[NH2+][C@H](C)[C@@H](O)C1=CC=CC=C1	36	166	222	1	2	1	12	0.30	0.5
(1R,2R)-pseudoephedrine[1]	C[NH2+][C@H](C)[C@H](O)C1=CC=CC=C1	36	166	222	1	2	1	12	0.30	0.5
(1S,2S)-pseudoephedrine[1]	C[NH2+][C@@H](C)[C@@H](O)C1=CC=CC=C1	36	166	222	1	2	1	12	0.30	0.5

[1]ammonium form

As can be seen in Table 2, the numerical properties produced from the 4 ephedrine and pseudoephedrine stereoisomers are the same. However, the values of the association constant with β-CD are different: -10.8, -10.6, -10.5, -11.3 kJ/mol, respectively, in the same conditions (pH = 6.9 and temperature is 24.9°C) [13]. RDKit and PubChem are not capable of generating/calculating specific physico-chemical properties such as melting points, boiling points, or optical activity to discriminate these molecules. The only way to discriminate these molecules is to check the IsomericSMILES property.

Unfortunately, machine learning algorithms often require numerical data. Therefore, it is needed to translate the textual IsomericSMILES property into some numerical properties. The word2vec[20] approach is commonly used to project textual values into a vector space (a tuple of numerical values). Word2vec was designed to convert words from a text into vectors in such a way that words with similar

semantics in the text have close spatial positions. We designed a similar approach that we called Iso2vec, which aims at representing molecules as vectors in such a way that similar molecules have close spatial positions. In this approach, not published yet, the words are the different characters of the IsomericSMILES (such as 'C', 'O', 'N', 'H', '1', '2', '3', etc. but also '(', ')', '#', '=', '[', ']', '+' and '@'), and the text corresponds to the IsomericSMILES itself. Iso2vec produces a vector of 10 numerical values for each molecule's IsomericSMILES representation.

Iso2vec just like word2vec is a prediction-based model. Therefore, the vector returned by this model for each molecule is a prediction, and this prediction depends on the data the model was trained on. The number of dimensions needed to represent the words or molecules depends on the training data as well. The training data for building Iso2vec consists of the 1789 IsomericSMILES values of both host and guest molecules in the dataset.

Table 3 shows the vectors produced by Iso2vec for the molecules in Table 2.

Table 3

Vectors generated by Iso2vec for the molecules from Table 2.

IsomericSMILES	0	1	2	3	4	5	6	7	8	9
CC#N	-0.27	-0.29	-0.48	0.78	0.43	-0.34	0.27	0.22	0.36	-0.16
CCCCCC[NH3+]	-0.26	0.36	-0.08	0.99	1.46	0.13	0.88	1.49	1.95	0.04
CC(C)C(=O)O	0.63	0.00	-0.9	1.03	0.61	-0.26	1.01	1.12	1.16	-0.03
C[NH2+][C@@H](C)[C@H](O)C1=CC=CC=C1	0.28	-0.22	-0.11	0.39	-0.05	-0.06	0.09	0.12	0.08	-0.08
C[NH2+][C@H](C)[C@@H](O)C1=CC=CC=C1	0.27	-0.22	-0.11	0.38	-0.06	-0.06	0.09	0.14	0.1	-0.08
C[NH2+][C@H](C)[C@H](O)C1=CC=CC=C1	0.28	-0.24	-0.11	0.35	-0.1	-0.07	0.09	0.14	0.1	-0.06
C[NH2+][C@@H](C)[C@@H](O)C1=CC=CC=C1	0.28	-0.20	-0.12	0.42	-0.02	-0.05	0.09	0.12	0.08	-0.09

In Table 3, the vectors produced for (1R,2S)-ephedrine, (1S,2R)-ephedrine, (1R,2R)-pseudoephedrine, and (1S,2S)-pseudoephedrine molecules are close to each other, but different. Those new numerical values allow us to distinguish the four molecules while the previous numerical values could not.

We finally obtain for each pair of CD and guest molecules, on top of the 6 original values: 2 identifiers, 7+9 physicochemical properties, and 10*2 numerical values thanks to Iso2vec. This forms a total of 44 properties, among which 40 consist in numerical values, making them directly usable with most machine learning approaches.

3. Data, value and validation

The PDF files are processed as images. For this reason, dividing the tables into columns or even recognizing some values in the tables could be difficult. For example, the value "4,4•-Dibromobiphenyl" in a PDF table [15] has been recognized as "4,4N-Dibromobiphenyl" by our script and the query to PubChem would return "4,4'-Dibromobiphenyl". In such case, the data is ignored. In another PDF file [16], the tables were not correctly read and therefore the values of temperature, K and ΔG values were mixed. Such errors and similar ones have been fixed manually.

There are several cases for which the data correctly gathered from various sources could not be included in the dataset.

First, the extracted values that are significantly different from the rest of the collected data are called outliers. The original articles from which the outlier values have been observed were checked and either validated or discarded. In some articles, the captions of some tables mention that some values may be erroneous. For example, take α-CD and indole pair for which ΔG is -44.3 ± 0.3 kJ/mol (log K = 7.8 ± 0.1) under normal conditions. This value is considered an outlier. Indeed, in the original article [27] where this data was gathered, there is a note “possible systematic errors in the calculations at values of log K > 5”. Such proven outliers were removed from the dataset.

Second, the same CD-guest pair can be found in several sources. If the association constant (or Gibbs free energy) is the same, only one occurrence of the pair is used with the original reference. If the different sources do not agree on a value, the one with the most accurate technique is selected. Remaining occurrences are discarded.

Third, in accordance with [11], only the data from Binding DB with pH values between 6.9 and 7.4 and temperature values between 14.5 and 30.1°C were included [13].

Fourth, some data were obtained in other solvents than H_2O [16]. In this case, these values were not used because incomparable with the others.

Furthermore, on data retrieved from [15], the guest molecule may be adjusted at the light of the charge (e.g., *acetic acid* with charge -1 actually means *acetate*). In that case, the name of the guest molecule is

unchanged to facilitate later data extraction checking but the CID is updated to the correct molecule. When the charged molecule could not be found on PubChem (no CID), we generated the IsomericSMILES manually from the uncharged molecule IsomericSMILES.

To summarize, the total data in the raw data file was equal to 3754 CD-guest pairs (the pie chart with a large radius; figure 2A). After having eliminated some data as described above, the curated final dataset file contains 3459 CD-guest pairs (the pie chart with a small radius, the light-colored parts show the eliminated data, and the dark parts show the data distribution in the final file; figure 2B). So 92% of the original data were retained. Finally, for most guest molecules, their name was used to retrieve from PubChem an IsomericSMILES and a PubChem identifier (CID). Such CID was used from time to time to check that the name, IsomericSMILES and CID still match (since the database evolves, the name evolves, and such condition may not always hold). We removed from the dataset the CD-guest pairs for which such property did not hold.

The data files (raw data and final enriched data) and the script files are available at URL https://doi.org/10.5281/zenodo.7575539 and https://doi.org/10.5281/zenodo.7575579, respectively. The dataset is released under the Open Data Commons Open Database License v1.0. The script files are released under the BSD 3-Clause.

To conclude, this curated dataset can be used to find existing or predict new association constant between a CD and a guest. Machine learning based prediction is of great interest because no analytical method, no toxic solvent or no sample are necessary. This database is easily accessible under a universal format. This database could interest chemists, analytical researchers or pharmaceuticals, but also machine learning specialists because it becomes yet another benchmark for them. The ultimate goal would be that this database serves as a basis for an enriched database collaboratively maintained by all research groups with knowledge of association constant values between various cyclodextrin and guest pairs.

References

[1] G. Crini, Review: A History of Cyclodextrins, Chem. Rev. 114 (2014) 10940–10975. https://doi.org/10.1021/cr500081p.

[2] A.R. Hedges, Industrial Applications of Cyclodextrins, Chem. Rev. 98 (1998) 2035–2044. https://doi.org/10.1021/cr970014w.

[3] H.-J. Schneider, F. Hacket, V. Rüdiger, H. Ikeda, NMR Studies of Cyclodextrins and Cyclodextrin Complexes, Chem. Rev. 98 (1998) 1755–1786. https://doi.org/10.1021/cr970019t.

[4] L. Hernández-García, A. Rojas-Hernández, A. Galano, Mangiferin/β-cyclodextrin complex: determination of the Inclusion constant in aqueous solution by Higuchi–Connors method and molecular absorption and photoluminescence UV spectroscopies at pH 3.4, Chem. Pap. 76 (2022) 7123–7132. https://doi.org/10.1007/s11696-022-02381-z.

[5] H. Aki, T. Niiya, Y. Iwase, M. Yamamoto, Calorimetry to Evaluate Inclusion Mechanism in the Complexation Between 2-hydroxypropyl-β-cyclodextrin and Barbiturates in Aqueous Solution, J. Therm. Anal. Calorim. 64 (2001) 713–719. https://doi.org/10.1023/A:1011592327676.

[6] Z.J. Baum, X. Yu, P.Y. Ayala, Y. Zhao, S.P. Watkins, Q. Zhou, Artificial Intelligence in Chemistry: Current Trends and Future Directions, J. Chem. Inf. Model. 61 (2021) 3197–3212. https://doi.org/10.1021/acs.jcim.1c00619.

[7] T.F.G.G. Cova, A.A.C.C. Pais, Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns, Front. Chem. 7 (2019) 809. https://doi.org/10.3389/fchem.2019.00809.

[8] A. Merzlikine, Y.A. Abramov, S.J. Kowsz, V.H. Thomas, T. Mano, Development of machine learning models of β-cyclodextrin and sulfobutylether-β-cyclodextrin complexation free energies, Int. J. Pharm. 418 (2011) 207–216. https://doi.org/10.1016/j.ijpharm.2011.03.065.

[9] Q. Zhao, Z. Ye, Y. Su, D. Ouyang, Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling

techniques, Acta Pharm. Sin. B. 9 (2019) 1241–1252. https://doi.org/10.1016/j.apsb.2019.04.004.

[10] M. Mizera, E.N. Muratov, V.M. Alves, A. Tropsha, J. Cielecka-Piontek, Computer-Aided Discovery of New Solubility-Enhancing Drug Delivery System, Biomolecules. 10 (2020) 913. https://doi.org/10.3390/biom10060913.

[11] R.M. Carvalho, I.G.L. Rosa, D.E.B. Gomes, P.V.Z.C. Goliatt, L. Goliatt, Gaussian processes regression for cyclodextrin host-guest binding prediction, J. Incl. Phenom. Macrocycl. Chem. 101 (2021) 149–159. https://doi.org/10.1007/s10847-021-01092-4.

[12] E. Hazai, I. Hazai, L. Demko, S. Kovacs, D. Malik, P. Akli, P. Hari, J. Szeman, E. Fenyvesi, E. Benes, L. Szente, Z. Bikadi, Cyclodextrin knowledgebase a web-based service managing CD-ligand complexation data, J. Comput. Aided Mol. Des. 24 (2010) 713–717. https://doi.org/10.1007/s10822-010-9368-y.

[13] T. Liu, Y. Lin, X. Wen, R.N. Jorissen, M.K. Gilson, BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities, Nucleic Acids Res. 35 (2007) D198–D201. https://doi.org/10.1093/nar/gkl999.

[14] J.L. Lach, J. Cohen, Interaction of Pharmaceuticals with Schardinger Dextrins II, J. Pharm. Sci. 52 (1963) 137–142. https://doi.org/10.1002/jps.2600520207.

[15] K.A. Connors, Population Characteristics of Cyclodextrin Complex Stabilities in Aqueous Solution, J. Pharm. Sci. 84 (1995) 843–848. https://doi.org/10.1002/jps.2600840712.

[16] M.V. Rekharsky, Y. Inoue, Complexation Thermodynamics of Cyclodextrins, Chem. Rev. 98 (1998) 1875–1918. https://doi.org/10.1021/cr970015o.

[17] T. Suzuki, A Nonlinear Group Contribution Method for Predicting the Free Energies of Inclusion Complexation of Organic Molecules with α- and β-Cyclodextrins, J. Chem. Inf. Comput. Sci. 41 (2001) 1266–1273. https://doi.org/10.1021/ci010295f.

[18] A. Lantz, M. Rodriguez, S. Wetterer, D. Armstrong, Estimation of association constants between oral malodor components and various native and derivatized cyclodextrins, Anal. Chim. Acta. 557 (2006) 184–190. https://doi.org/10.1016/j.aca.2005.10.005.

[19] M. Kfoury, L. Auezova, H. greige-gerges, S. Fourmentin, Encapsulation in cyclodextrins to widen the applications of essential oils, Environ. Chem. Lett. 17 (2018). https://doi.org/10.1007/s10311-018-0783-y.

[20] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, ArXiv13013781 Cs. (2013). http://arxiv.org/abs/1301.3781 (accessed December 13, 2021).

[21] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2023 update, Nucleic Acids Res. 51 (2023) D1373–D1380. https://doi.org/10.1093/nar/gkac956.

[22] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Model. 28 (1988) 31–36. https://doi.org/10.1021/ci00057a005.

[23] P. Ertl, B. Rohde, P. Selzer, Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties, J. Med. Chem. 43 (2000) 3714–3717. https://doi.org/10.1021/jm000942e.

[24] S.H. Bertz, The first general index of molecular complexity, J. Am. Chem. Soc. 103 (1981) 3599–3601. https://doi.org/10.1021/ja00402a071.

[25] J.B. Hendrickson, P. Huang, A.G. Toczko, Molecular complexity: a simplified formula adapted to individual atoms, J. Chem. Inf. Comput. Sci. 27 (1987) 63–67. https://doi.org/10.1021/ci00054a004.

[26] S.A. Wildman, G.M. Crippen, Prediction of Physicochemical Parameters by Atomic Contributions, J. Chem. Inf. Comput. Sci. 39 (1999) 868–873. https://doi.org/10.1021/ci990307l.

[27] E.A. Lewis, L.D. Hansen, Thermodynamics of binding of guest molecules to α- and β-cyclodextrins, J. Chem. Soc. Perkin Trans. 2. (1973) 2081–2085. https://doi.org/10.1039/P29730002081.