BigSolDB: Solubility Dataset of Compounds in Organic Solvents and Water in a Wide Range of Temperatures

Lev Krasnov[1], Simon Mikhaylov[1], Prof. Maxim V. Fedorov[2], and Dr. Sergey Sosnin[3,*]

[1]Department of Chemistry, Lomonosov Moscow State University, 119991 Moscow GSP-1, 1-3 Leninskiye Gory, Russia
[2]Sirius University of Science and Technology, Olimpiysky ave. b.1, Sochi, 354000 , Russia
[3]Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, Josef-Holaubek-Platz 2, 1190 Vienna, Austria
[*]corresponding author(s): Dr. Sergey Sosnin (sergey.sosnin@univie.ac.at)

ABSTRACT

Solubility is crucial in organic chemistry and holds significant value in the field of medicinal chemistry. Employing computational and QSPR modeling for solubility estimation is favorable as it reduces experimental costs. However, high-quality experimental data is essential for training these QSPR models. In our study, we compiled a dataset consisting of 54,273 experimental solubility values within a temperature range of 243.15 to 403.15 K in various organic solvents and water. This dataset can be used as a reference for individual values or training solubility QSPR models. We conducted a statistical analysis and identified prevalent patterns in the data. Furthermore, we developed an interactive, parametric t-SNE-based tool to explore the chemical space of solutes. Utilizing this tool, we characterized common scaffolds in the dataset and demonstrated that the chemical space of solutes is extensive and diverse.

Background & Summary

Solubility is the property of a chemical solute dissolving in a solvent to form a homogeneous system. Solubility depends on the solvent used and the pressure and temperature at which it was measured[1]. While the solubility of solids in the majority of organic solvents plays an essential role in various chemical fields such as organic chemistry, environmental chemistry, organo(photo)catalysis, and petrochemical industry; the solubility data in specified solvents such as water or dimethyl sulfoxide (DMSO) is highly crucial for drug development[2–5]. For example, low solubility in water is attributed with poor bioavailability because compounds that are non-soluble in DMSO could not be used in automatized high-throughput screening (HTS). Generally, it makes them lost for experimental measurements[4].

There are two general ways to solubility modelling: bottom-up and top-down. Bottom-up approaches are primarily based on molecular simulations[6,7]. Top-down approaches are based on thermodynamic models and quantitative structure property relationship (QSPR) methods[8]. Nowadays, QSPR methods for solubility predictions heavily use machine learning as statistical backbone.[9–11] Machine learning based methods demonstrate high-performance, but rely entirely on the quality and amount of training data[12]. Providing high-quality data is of great importance for further development of top-down approaches.

There are many possible modeling endpoints related to solubility. Solubility can be measured in various solvents and at different temperatures. The interest of pharmaceutical researchers is biased towards the solubility in water and DMSO under normal conditions. However, for organic chemistry, many other solvents can be at the same level of importance. During synthesis and for the purification and crystallization of compounds, one needs to know solubility of compounds in various organic solvents. Our observations indicates that there is lack of solubility data measured in organic solvents, rather than water and DMSO[13,14]. Because solubility is a temperature-dependent property, the solubility data for a number of temperatures is also important. Several solubility datasets currently exist covering a wide range of temperatures and solvents, but the number of measured compounds is not enough for efficient training of machine learning models .[15–17]

The aim of this work is to create a large dataset of organic compounds solubility values to supplement available data for further data-driven model development. For this purpose we searched for and collected data from 771 unique articles from 3 previously selected journals (Figure 1). As a result, organic compounds solubilities in 138 different solvents were collected. This dataset includes solubility values for 839 unique organic compounds under temperatures from 243,15 K to 403,15 K at atmospheric pressure. The complete dataset contains 54273 solubility data records. We named this dataset BigSolDB.

We present an open-access dataset of solubility in organic solvents and water in a wide range of temperatures as the main outcome of our research. We expect it to serve a broad community as a reference dataset for the bench-marking of new experimental modeling results and also as the data resource for future machine-learning investigations in the chemoinformatics field, for example, to improve the prediction capability of future models.

Humans prefer the visual examination of large volumes of data. Recently, we proposed the parametric t-SNE method as a universal tool for visualizing chemical data[18,19]. In this research, we provide an interactive web-based tool to visualize the chemical space of solutes. Using our tool, chemists can explore our dataset and see the solubility charts for each compound individually. Additionally, we provided the description of significant scaffolds in BigSolDB dataset.

Methods

Data collection

The dataset construction was conducted in four steps: (1) choosing appropriate journals, (2) getting data from Crossref API and filtering papers, (3) extracting data from papers, (4) cleaning and preprocessing data. Our steps are outlined in Figure 1.

Journals Selection

In the first step, we selected three journals containing solubility data: *Journal of Chemical & Engineering Data* (ISSN: 1520-5134), *The Journal of Chemical Thermodynamics* (ISSN: 0021-9614), and *Fluid Phase Equilibria* (ISSN: 0378-3812). These journals focused on publishing compounds' physicochemical data. The motivation behind choosing these journals was that they systematically publish compounds' solubility data for several solvents at various temperatures.

Getting data from Crossref API and filtering papers

In the second step, we searched for the International Standard Serial Number (ISSN) for each journal and, using the Crossref API (`https://api.crossref.org/`) and obtained a list of Digital Objects Identificators (DOIs) and titles for each paper.

- *Journal of Chemical & Engineering Data*: (ISSN: 1520-5134)
- *The Journal of Chemical Thermodynamics*: (ISSN: 0021-9614)
- *Fluid Phase Equilibria*: (ISSN: 0378-3812)

Using the Crossref bibliographic data, we searched for "solub" subword in titles to pick only papers with solubility data. We performed manual processing to remove papers that contain "solub" in their titles but do not have the solubility data we needed. At the final stage, we had 771 papers ready to further processing.

Extracting data from papers

We recognized and converted solubility data into a machine-readable format manually. We focused only on mole fraction solubility under atmospheric pressure. We used OPSIN open-source cheminformatics software (`https://opsin.ch.cam.ac.uk/`) to convert IUPAC names of compounds into SMILES representation.

Cleaning and preprocessing data

In the chemical literature, solvents can be represented in various ways, which poses a significant challenge for dataset preparation. To address this issue, we standardized the solvents to eliminate ambiguity. For instance, we converted terms like *"ethyl alcohol"* and *"methylcarbinol"* to *"ethanol"* while *"2-butanol"*, *"butan-2-ol"*, and *"sec-butyl alcohol"* were transformed into *"sec-butanol"*. Subsequently, we employed the OSPIN system to convert the standardized solvent names. Lastly, we utilized the open-source cheminformatics software RDKit (`https://www.rdkit.org`) to perform SMILES canonization for each solute compound.

Parametric t-SNE visualization

To make exploring our database easier for the researchers, we prepared an interactive web-based tool to visualize the solutes' chemical space (Figure 2). Our tool helps navigate the 2D scatter plot leveraging several standard operations (zoom areas, pick compounds, etc.). Also, one can filter compounds by a solvent of interest. The projection is based on the parametric t-SNE method. The parametric t-SNE model is an artificial neural network that projects molecular structures to 2D (X and Y) coordinates, keeping the local similarity between molecules. We used the pretrained model from our previous work[18]. This model was trained on complete ChEMBL data and provided good discriminating ability between organic scaffolds. The advantage of the parametric t-SNE over the vanilla t-SNE is the determinacy of the model. Because a pretrained model is an artificial neural network with fixed coefficients, it projects new compounds to known-in-advance regions of chemical space. Utilizing this feature, one can compare several datasets together, using one of them as the reference dataset.

Data Records

In this research we created BigSolDB: a new broad solubility dataset of a wide range of organic compounds. We provide it as a comma-separated values (csv) file. BigSolDB contains 54273 individual solubility values, 830 unique molecules and 138 individual solvents for the temperature range from 243.15 to 403.15K at atmospheric pressure. The dataset is freely accessible in Zenodo data sharing platform[20].

The file has five columns:

- SMILES — SMILES representation of a dissolved compound
- T,K — Temperature in Kelvin
- Solubility — an experimental solubility value (mole fraction)
- Solvent — a name of the solvent
- SMILES_Solvent — SMILES representation of the solvent
- Source — a data source for the record

Our interactive visualisation tool is freely accessible: `http://bigsoldbapp.cheminfo.space`. The project page is located: `https://bigsoldb.cheminfo.space`.

Data Analysis

One can see that top-8 solvents (ethanol, methanol, isopropanol, ethyl acetate, acetone, n-propanol, water, n-butanol, acetonitrile) cover 57% of data measurements in total, wherein water (3168 records, 5.8%) is only at the seventh place. Probably, this reflex the most common solvents that are using in organic chemistry. The similar picture we have for the individual compounds. The majority of experiments was performed in alcohols, not in water. However, this bias seems undesirable for the medicinal chemistry, because, the water solubility for drug candidates is of greater importance.

Following The United States Pharmacopeia (USP, a nonprofit organization committed to establishing standards for medicines, food, dietary supplements, etc.) approach, compounds can be divided into seven categories with respect to their solubilities ratios[21]. We converted the molar fractions into molar ratios under fixed temperature (293.15 K) and attributed them to one of seven USP classes (Table 2). The pie chart demonstrating the distribution of USP classes is given in Figure 14. One can see that six out of seven classes are well-represented in the dataset, the exception is "Very soluble" class which is presented only by 39 records (0.98% of total records at 293.15 K).

The data distribution across solvents was presented as a set of barplots, where each bar corresponds to the number of data rows (Figure 3, Figure 4) or the number of individual compounds (Figure 5, Figure 6) in the dataset. The top 50 of the most popular solvents were analysed separately (Figure 3, Figure 5).

The values of mole fraction solubilities shown as the histogram in Figure 7 are disturbed from $5.5 * 10^{-11}$ to $9.506 * 10^{-1}$. Figure 8 shows the distribution of the number of solubility observations measured at various temperatures.

Top-9 compounds that have the largest number of records are depicted in Figure 12. One can see, that 8 out of 9 of these compounds contain several functional groups with an oxygen atom (caboxyl-, nitro-, ester-, etc.).

Temperature coverage of the solubilities measurements in the 30 most popular solvents was presented in Figure 9. Right heat bar shows the counts of data rows for solubilities under corresponding temperatures. Temperatures to show were chosen according to the histogram in Figure 8. This figure shows the results for the temperature range with the largest number of records (from 278.15 to 323.15 K) in the dataset.

Also, data coverage was represented as a heatmap in Figure 10. We chose SMILES to show as the top 50 of the mostly described compounds in the dataset. And we analysed their solubilities in the 30 of the most popular solvents. Right heat bar represents to the data rows number of the solubilities of described compounds in shown solvents. As shown in the Figure 11, the majority of presented compounds occurs in the dataset only once. Because of this fact and the lack of the experimental errors information in the original articles, no statistical solubility data validation would be performed on the dataset.

We also compared BigSolDB with existing ones[15,17] and we present the results on Table 1. BigSolDB contains ten times more solubility values than previous datasets and three times more unique molecules in similar temperature range.

Chemical space of BigSolDB

To reveal the chemical space of BigSolDB, we prepared parametric t-SNE projections of BigSolDB and Tox21 dataset[22] (reference dataset) on the same plot (Figure 13). One should regard that the number of compounds in Tox21 dataset is larger than in BigSolDB. So, we had have to set a lower opacity value for Tox21 data points to prevent the visual suppression of BigSolDB

points. One can see that the overall distributions of chemical compounds are similar; both Tox21 and BigSolDB are broad and diverse. We highlighted and described several representative clusters on this map. The cluster (**A**) comprises "simple" aliphatic compounds. The cluster (**C**) is represented by aromatic compounds with the dominance of substituted benzenes. cluster (**H**) comprises of aliphatic heterocycles. The groups of steroids (**F**) and sugar derivatives (**G**) are well-recognisable. There are two clusters (**D**) and (**E**) which consists of two established clusters: N-phenyl-1-thia-3-azaspiro[5.5]undec-2-en-2-amine derivatives and Desethylenenorfloxacin derivatives. However, BigSolDB is notably enriched with sulfonamide derivatives (**B**) been compared with Tox21 dataset.

Usage Notes

The given dataset can be used both to view individual solubility values and to train machine learning models. To create fingerprints for machine learning models SMILES representations of dissolved compounds could be converted to an RDKit molecular object. The presented data is available under CC-BY 4.0 licence.

Code Availability

The Jupyter Notebook scripts for downloading data from Crossref API and plotting the figures can be found in https://github.com/Cemich/BigSolDB.

References

1. Boobier, S., Osbourn, A. & Mitchell, J. B. O. Can human experts predict solubility better than computers? *J. Cheminformatics* **9**, 63, 10.1186/s13321-017-0250-y (2017).
2. Delaney, J. S. Esol: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005, 10.1021/ci034243x (2004).
3. Sorkun, M. C., Khetan, A. & Er, S. Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Sci. Data* **6**, 143, 10.1038/s41597-019-0151-1 (2019).
4. Tetko, I. V. *et al.* Development of dimethyl sulfoxide solubility models using 163 000 molecules: Using a domain applicability metric to select more reliable predictions. *J. Chem. Inf. Model.* **53**, 1990–2000, 10.1021/ci400213d (2013).
5. Baybekov, S. *et al.* Dmso solubility assessment for fragment-based screening. *Molecules* **26**, 10.3390/molecules26133950 (2021).
6. Hossain, S., Kabedev, A., Parrow, A., Bergström, C. A. & Larsson, P. Molecular simulation as a computational pharmaceutics tool to predict drug solubility, solubilization processes and partitioning. *Eur. J. Pharm. Biopharm.* **137**, 46–55, https://doi.org/10.1016/j.ejpb.2019.02.007 (2019).
7. Bjelobrk, Z., Mendels, D., Karmakar, T., Parrinello, M. & Mazzotti, M. Solubility prediction of organic molecules with molecular dynamics simulations. *Cryst. Growth & Des.* **21**, 5198–5205, 10.1021/acs.cgd.1c00546 (2021). https://doi.org/10.1021/acs.cgd.1c00546.
8. Bouillot, B., Teychené, S. & Biscans, B. An evaluation of thermodynamic models for the prediction of drug and drug-like molecule solubility in organic solvents. *Fluid Phase Equilibria* **309**, 36–52, https://doi.org/10.1016/j.fluid.2011.06.032 (2011).
9. Kuentz, M. & Bergström, C. A. Synergistic computational modeling approaches as team players in the game of solubility predictions. *J. Pharm. Sci.* **110**, 22–34, https://doi.org/10.1016/j.xphs.2020.10.068 (2021).
10. Mullins, E., Liu, Y. A., Ghaderi, A. & Fast, S. D. Sigma profile database for predicting solid solubility in pure and mixed solvent mixtures for organic pharmacological compounds with cosmo-based thermodynamic methods. *Ind. & Eng. Chem. Res.* **47**, 1707–1725, 10.1021/ie0711022 (2008).
11. Ran, Y., Jain, N. & Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (gse). *J. Chem. Inf. Comput. Sci.* **41**, 1208–1217, 10.1021/ci010287z (2001).
12. Artrith, N. *et al.* Best practices in machine learning for chemistry. *Nat. Chem.* **13**, 505–508, 10.1038/s41557-021-00716-z (2021).
13. Abraham, M. H. *et al.* Prediction of solubility of drugs and other compounds in organic solvents. *J. Pharm. Sci.* **99**, 1500–1515, https://doi.org/10.1002/jps.21922 (2010).
14. Lim, H. & Jung, Y. Delfos: deep learning model for prediction of solvation free energies in generic organic solvents. *Chem. Sci.* **10**, 8306–8315, 10.1039/C9SC02452B (2019).

15. Ye, Z. & Ouyang, D. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. *J. Cheminformatics* **13**, 98, 10.1186/s13321-021-00575-3 (2021).

16. Boobier, S., Hose, D. R. J., Blacker, A. J. & Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **11**, 5753, 10.1038/s41467-020-19594-z (2020).

17. Vermeire, F. H., Chung, Y. & Green, W. H. Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. *J. Am. Chem. Soc.* **144**, 10785–10797, 10.1021/jacs.2c01768 (2022).

18. Karlov, D. S., Sosnin, S., Tetko, I. V. & Fedorov, M. V. Chemical space exploration guided by deep neural networks. *RSC Adv.* **9**, 5151–5157, 10.1039/C8RA10182E (2019).

19. Andronov, M., Fedorov, M. V. & Sosnin, S. Exploring chemical reaction space with reaction difference fingerprints and parametric t-sne. *ACS Omega* **6**, 30743–30751, 10.1021/acsomega.1c04778 (2021).

20. Krasnov, L., Mikhaylov, S., Fedorov, M. & Sosnin, S. Bigsoldb: solubility dataset of compounds in organic solvents and water in a wide range of temperatures. *zenodo* https://doi.org/10.5281/zenodo.6809668 (2022).

21. Convention, U. S. P. *USP35 NF30, 2012: U. S. Pharmacopoeia National Formulary*. The United States pharmacopeia (United States Pharmacopeial, 2011).

22. Huang, R. *et al.* Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front. Environ. Sci.* **3**, 10.3389/fenvs.2015.00085 (2016).

Acknowledgements

We used Figma (`https://www.figma.com/`) and Seaborn (`https://seaborn.pydata.org/`) to create all drawings in this paper. We used Dash (`https://dash.plotly.com/`) and plotly (`https://plotly.com/`) to create the visualisation tool.

Author contributions statement

L.K. and S.M. collected the dataset and analysed the results. S.S. implemented the tool for the visualisation of the dataset and provided the description of the chemical space. L.K., S.M. and S.S. wrote the draft of the manuscript, and M.V.F. revised it. S.S. supervised the project. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Figures & Tables

Dataset	Number of records	SMILES	Solvents	Temperature range	Sources
Dataset from[17]	4953	116	98	243.15-363.7	105
Dataset from[15]	5081	266	123	213.15-413.15	—
BigSolDB	54273	839	138	243.15-403.15	771

Table 1. Comparison of BigSolDB dataset with the previous ones.

Range of molar ratio	Class
0-1	Very soluble
1-10	Freely soluble
10-30	Soluble
30-100	Sparingly soluble
100-1000	Slightly soluble
1000-10000	Very slightly soluble
10000-∞	Practically insoluble

Table 2. USP solubility classes.

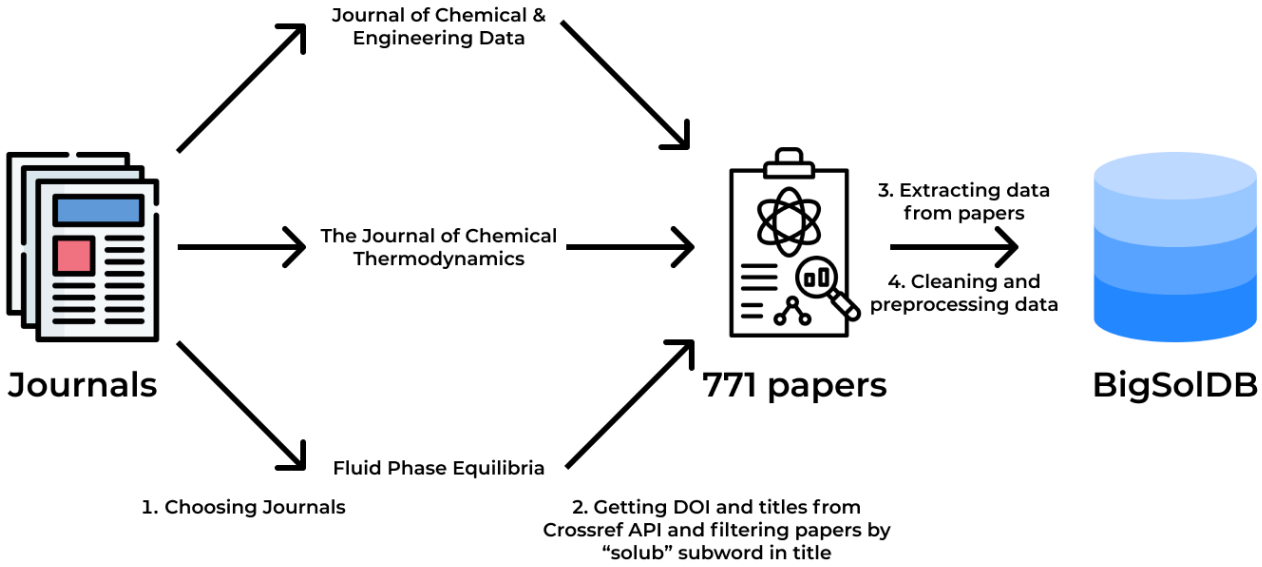

Figure 1. The process of creating a dataset.

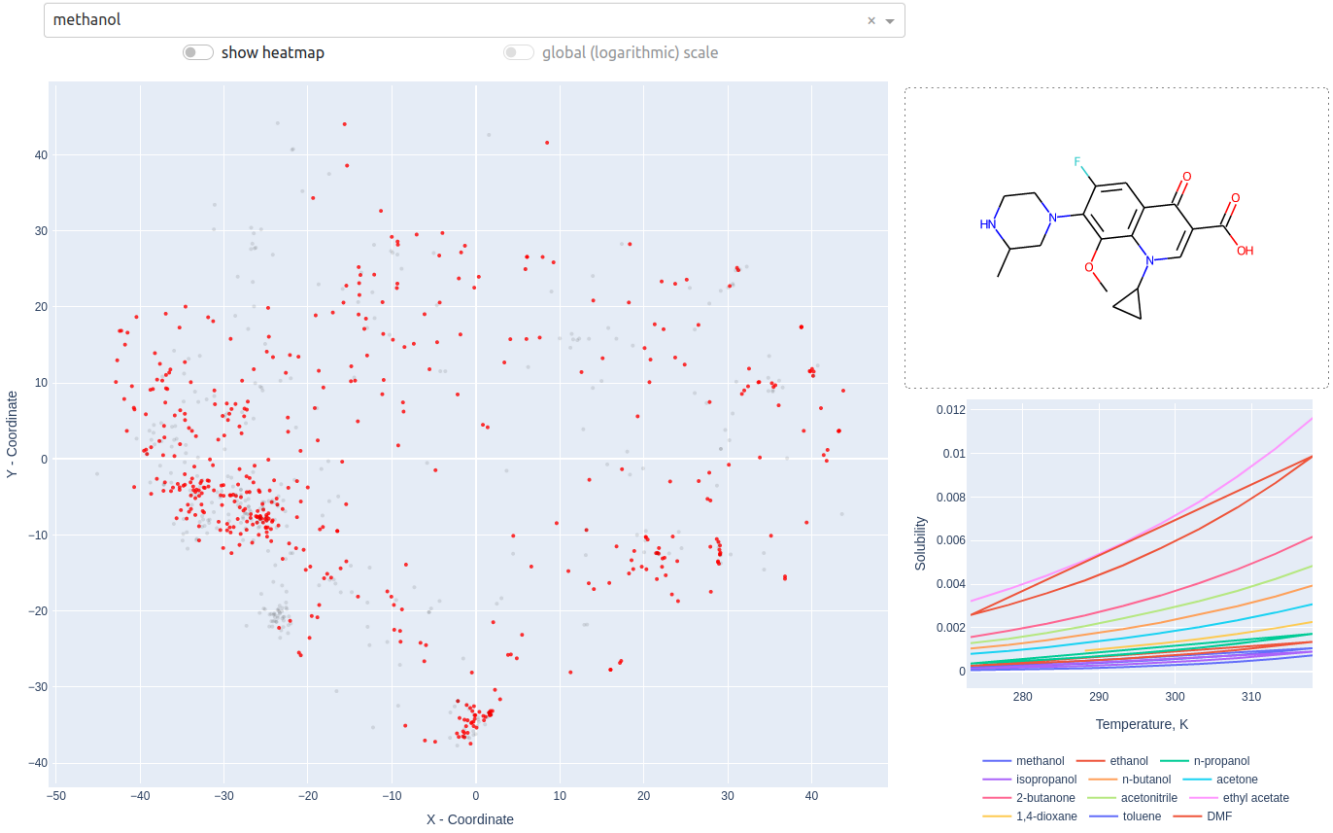

Figure 2. The demonstration of our solute chemical space visualization tool. On this figure, the particular solvent is selected "methanol", so the only compounds that have methanol solubility data are highlighted red. The rest of structures are paled grey.

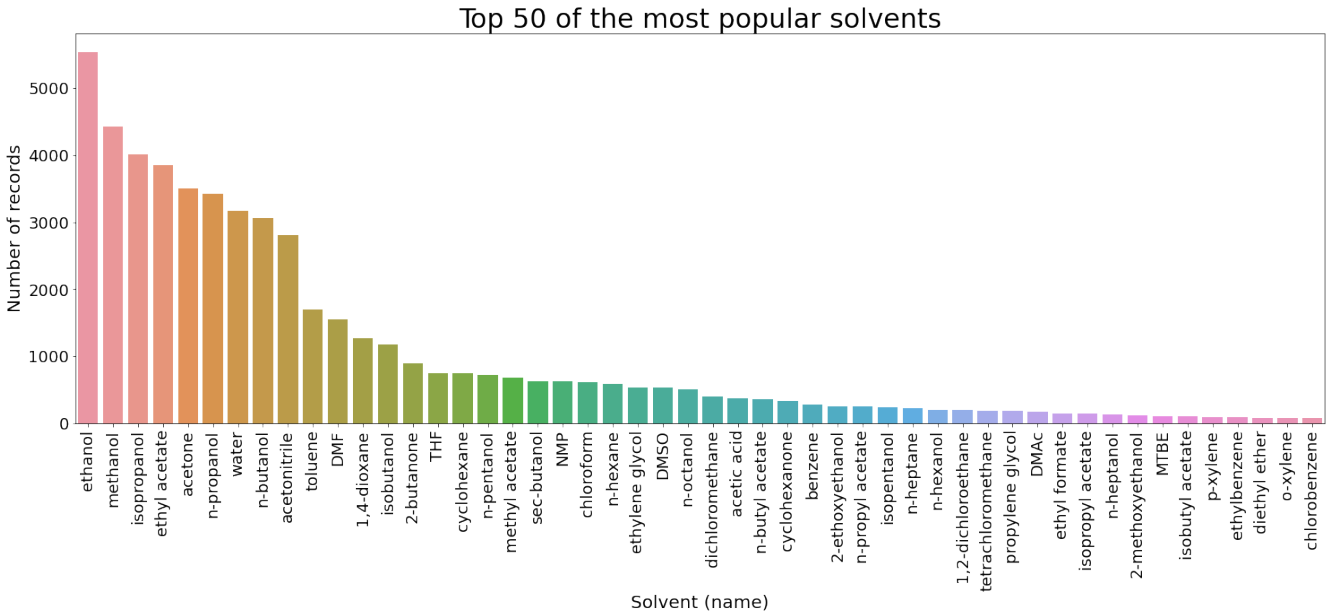

Figure 3. Barplot of the data rows number for the 50 most popular solvents.

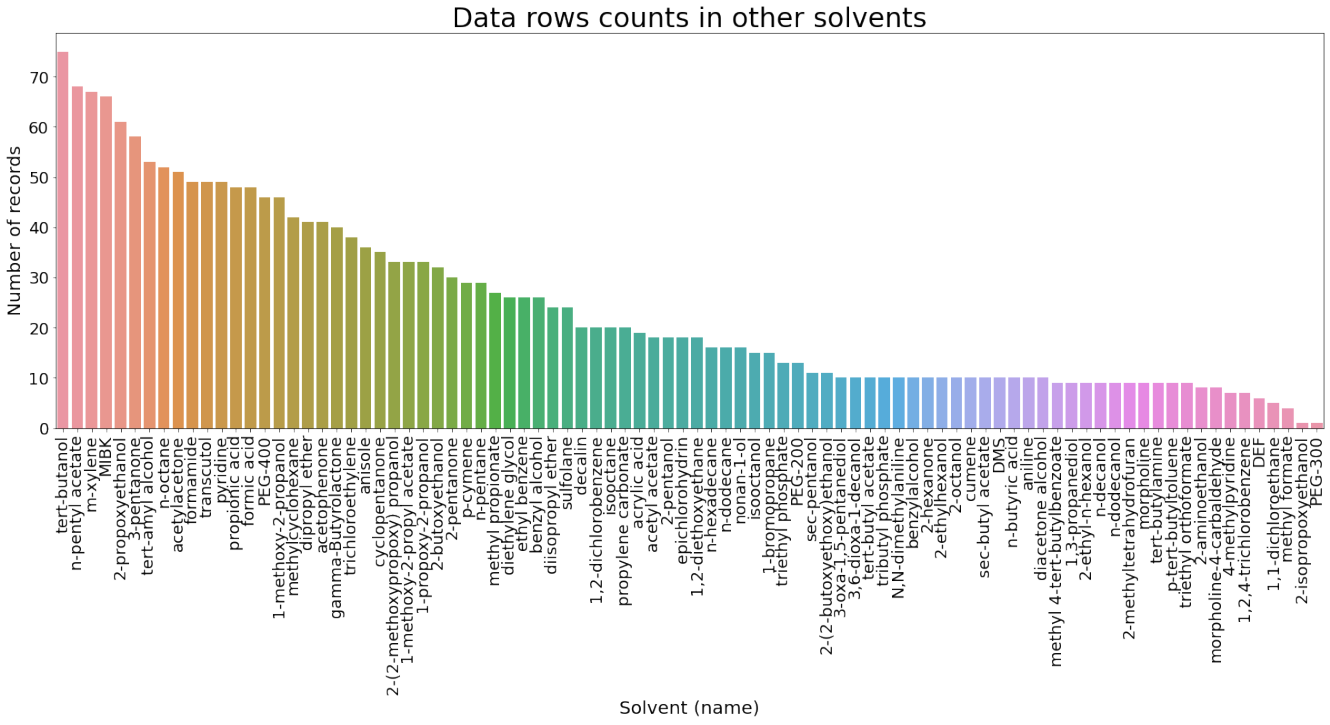

Figure 4. Barplot of the data rows number for the other solvents.

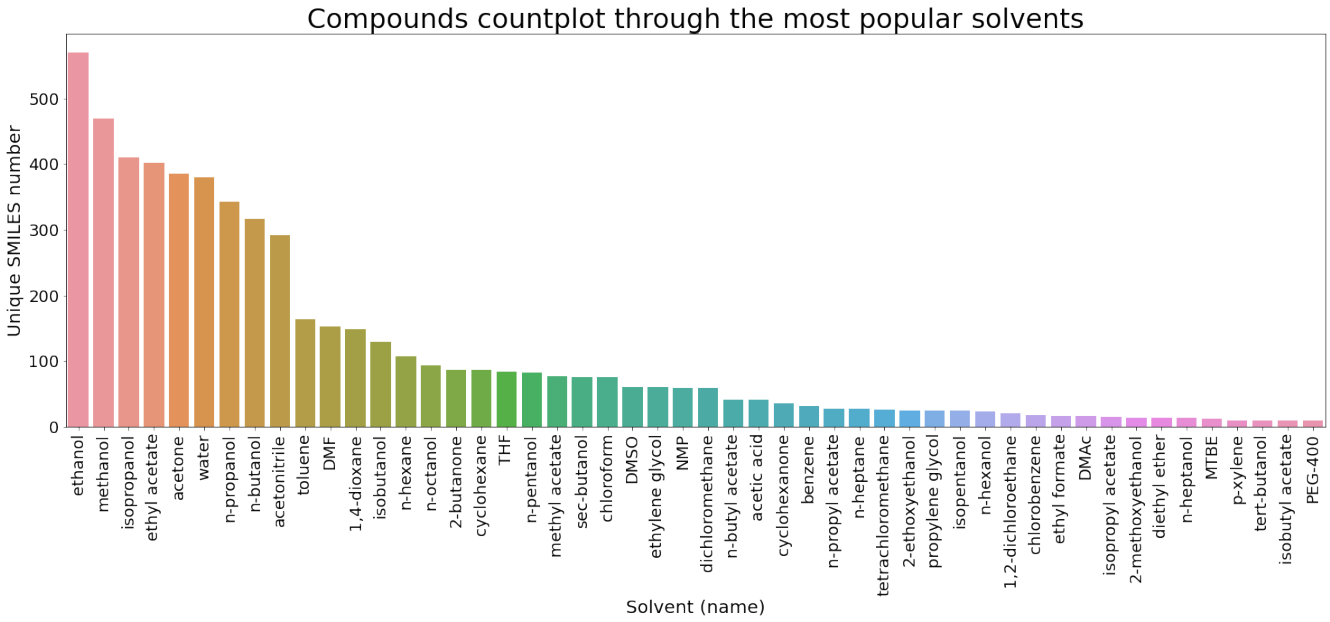

Figure 5. The number of compounds which were dissolved in the 50 most popular solvents.

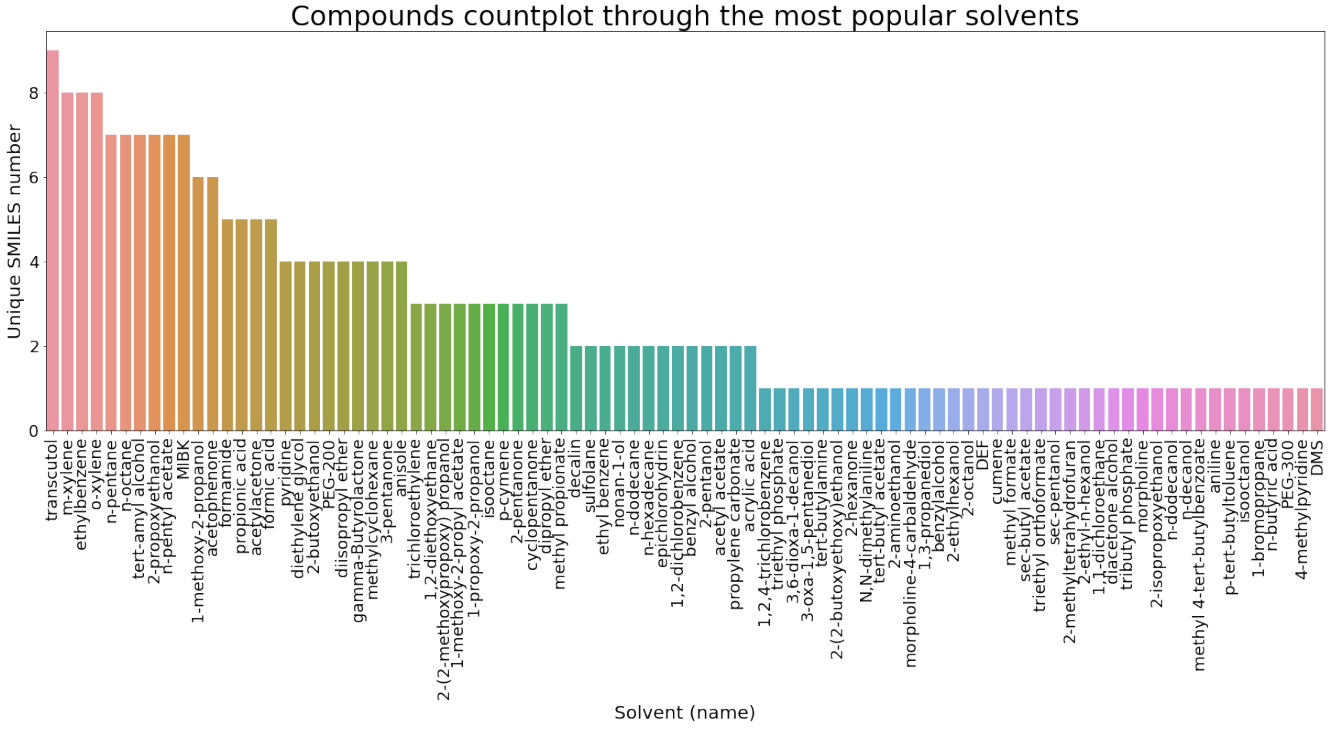

Figure 6. The numbers of compounds which were dissolved in the other solvents.

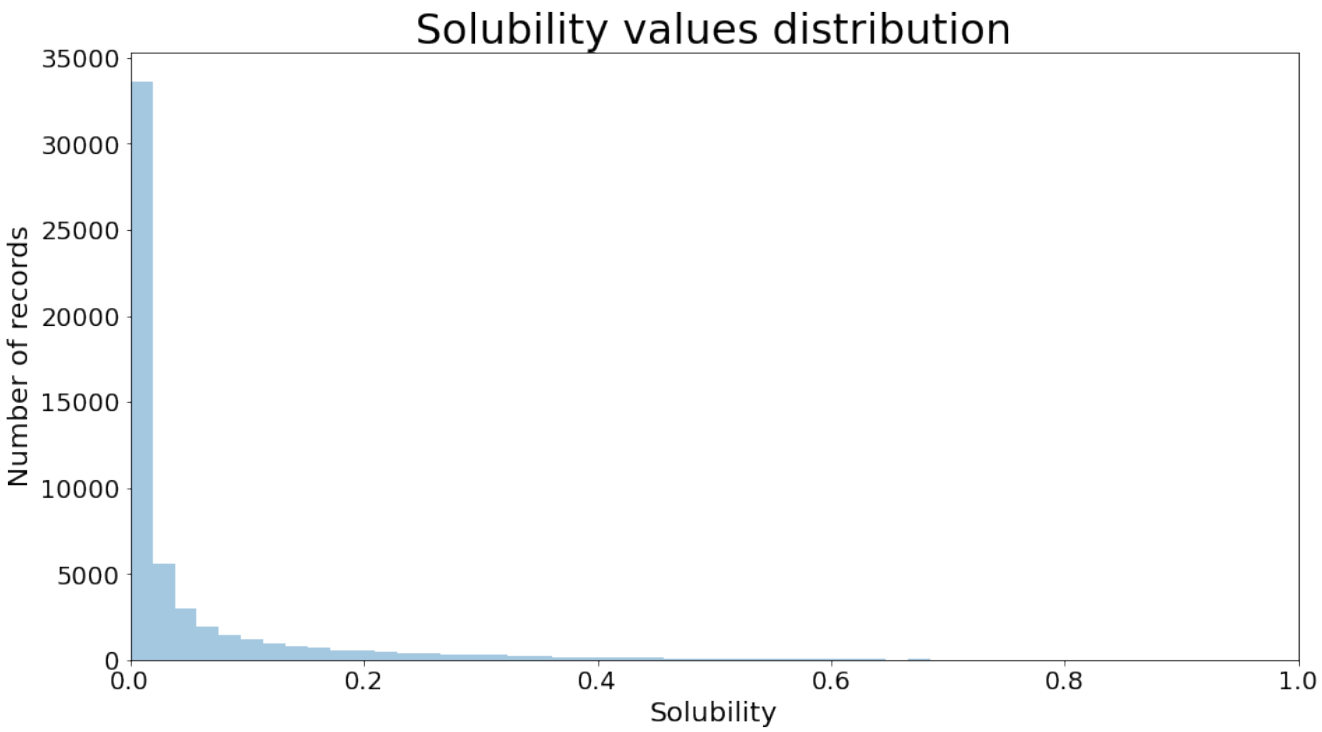

Figure 7. Histogram of all solubility values in the dataset.

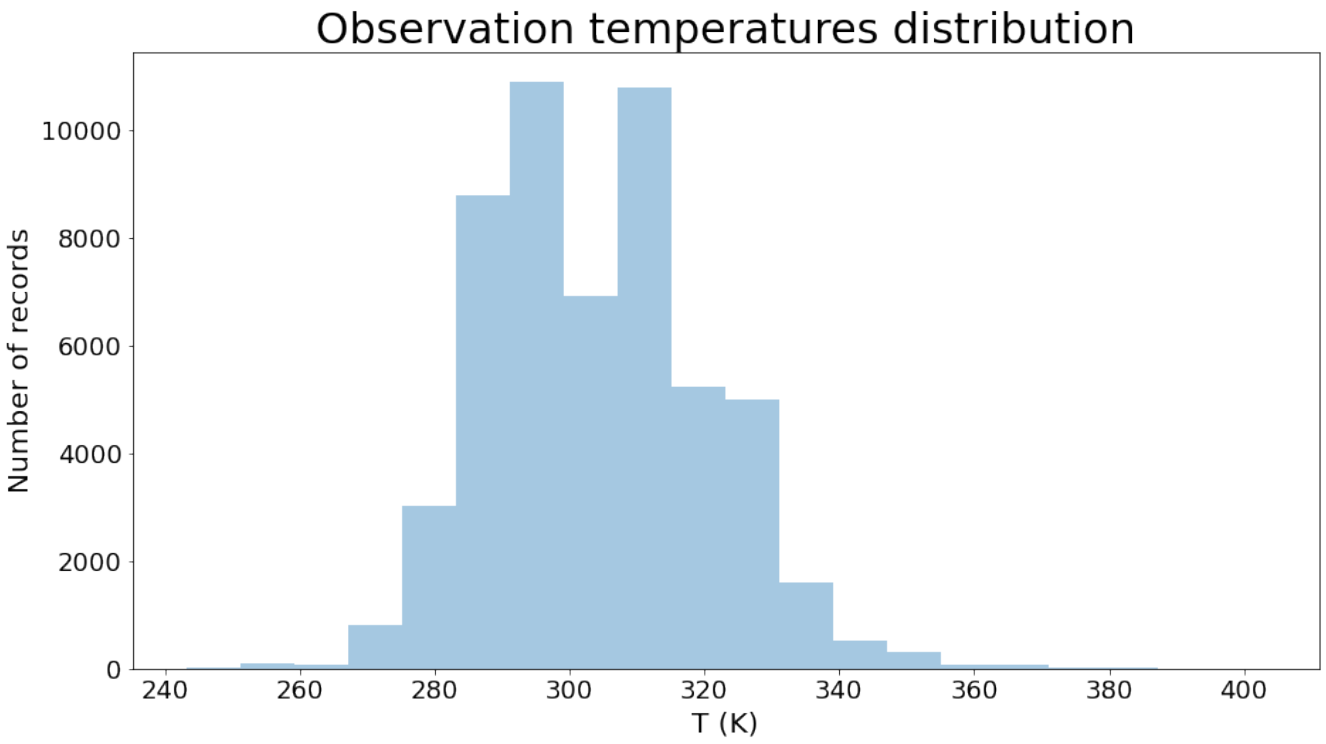

Figure 8. Histogram of all available temperatures of solubility observations.

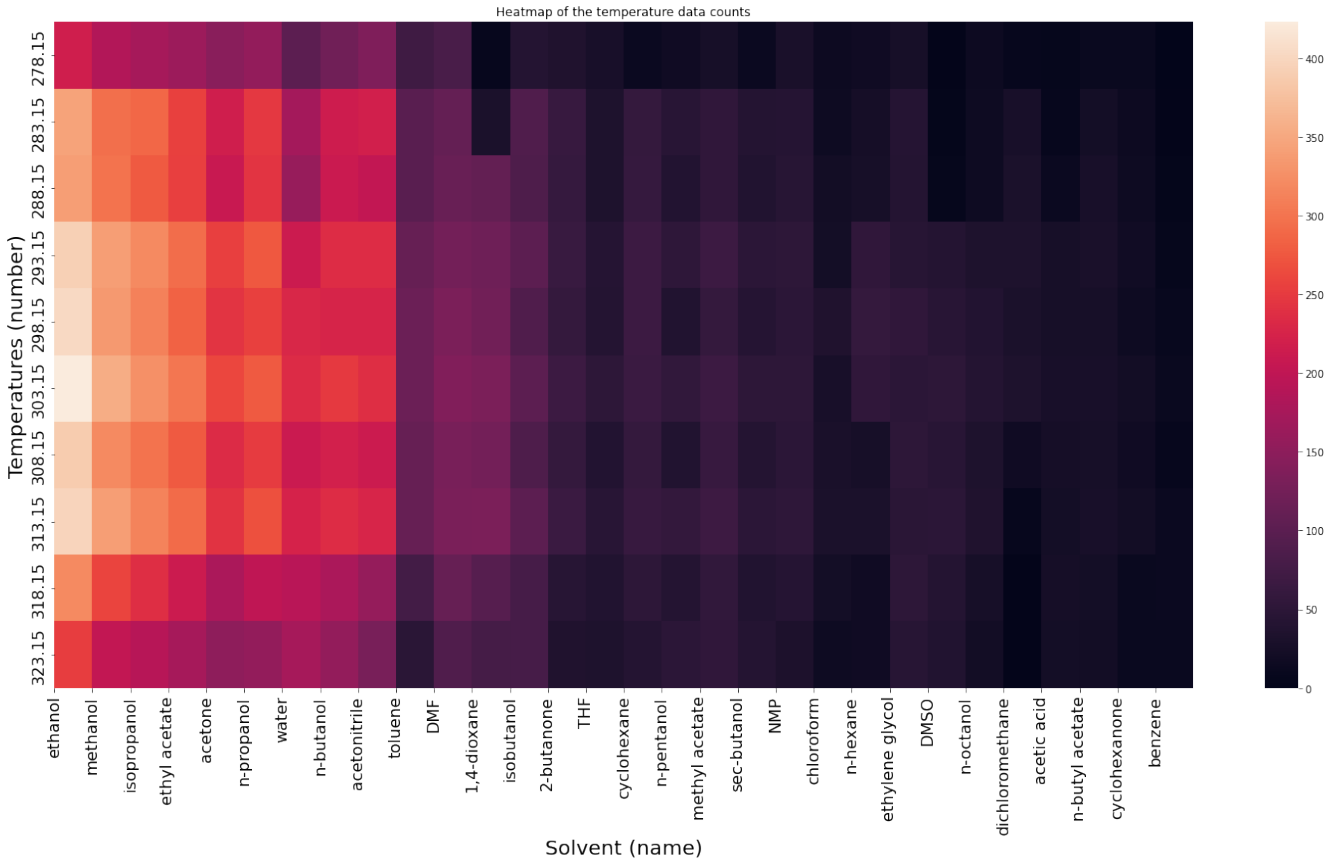

Figure 9. Heatmap of the number of solubilities measured in the 30 most popular solvents at the most popular temperatures.

Figure 10. Heatmap of the number of solubilities measured in the 30 most popular solvents for the compounds with the largest number of records.

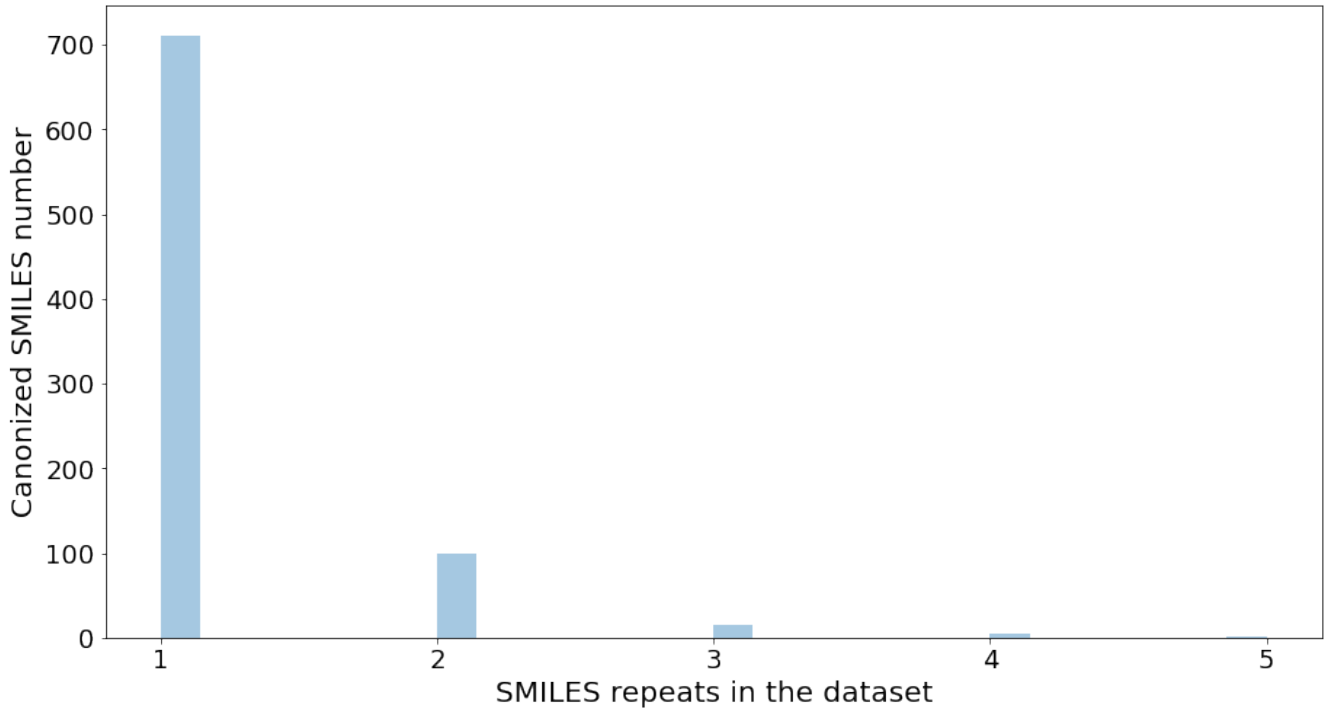

Figure 11. Countplot of the SMILES occurrence in the dataset.

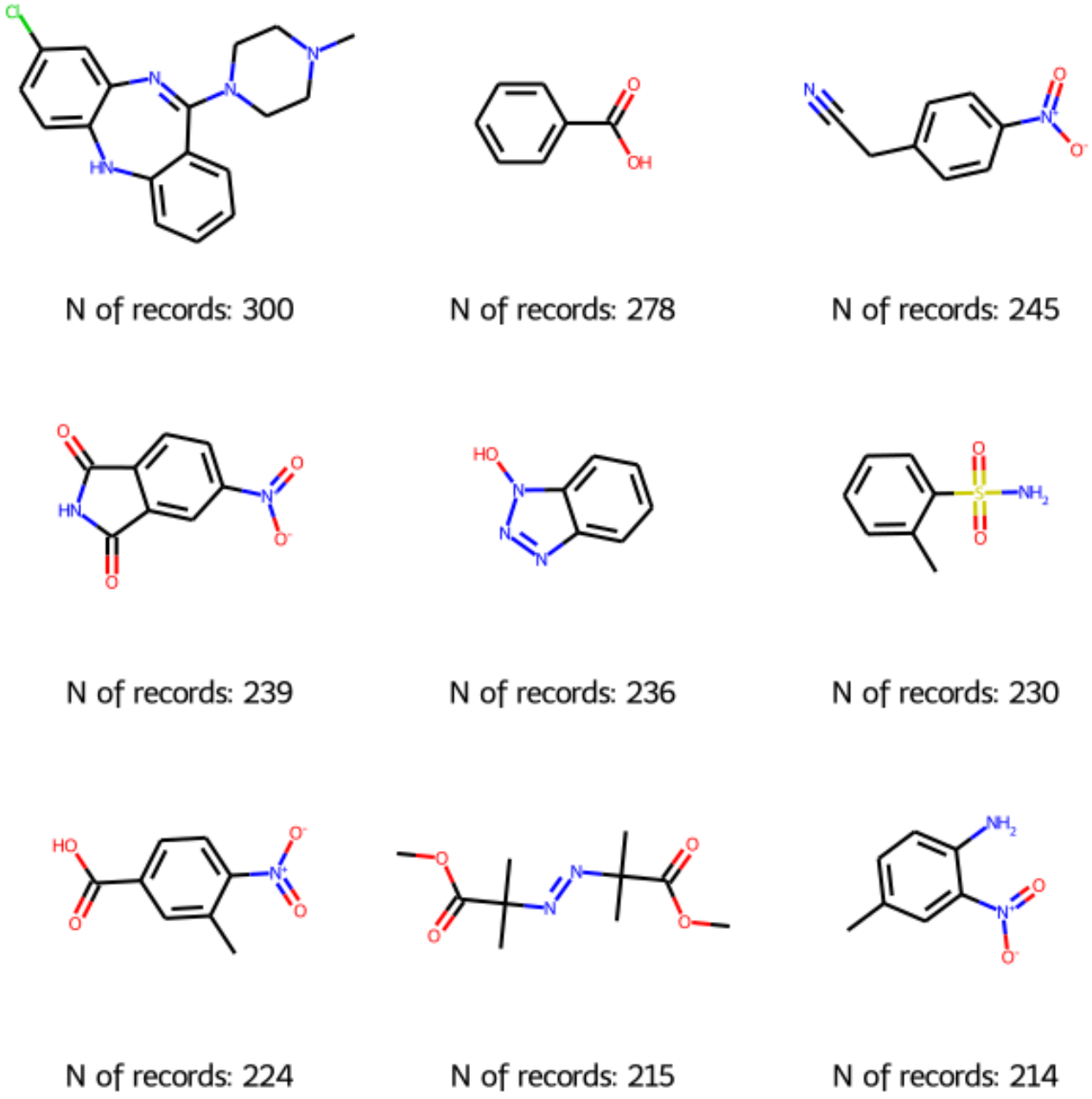

Figure 12. Top-9 of molecules sorted by the maximal number of records in the dataset.

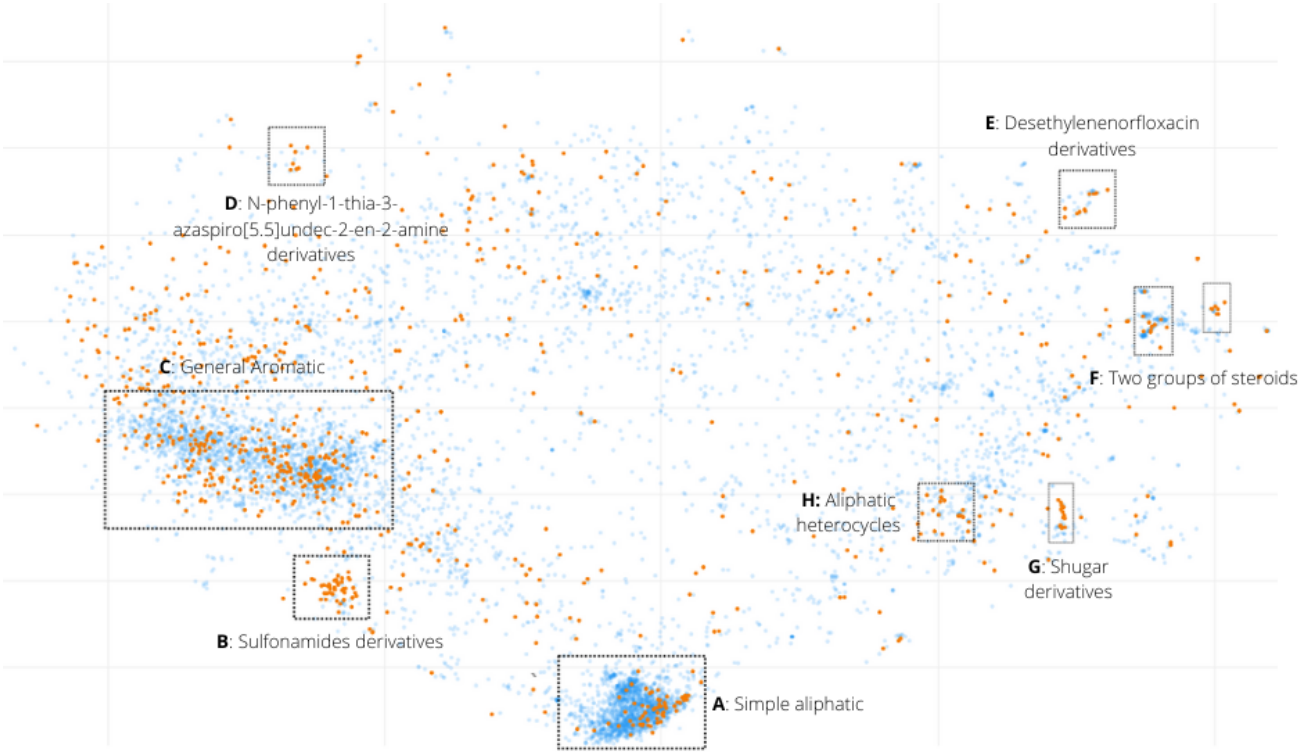

Figure 13. The comparison between Tox21 dataset and BigSolDB. Blue dots represent structures from Tox21 dataset, orange ones represent structures from BigSolDB

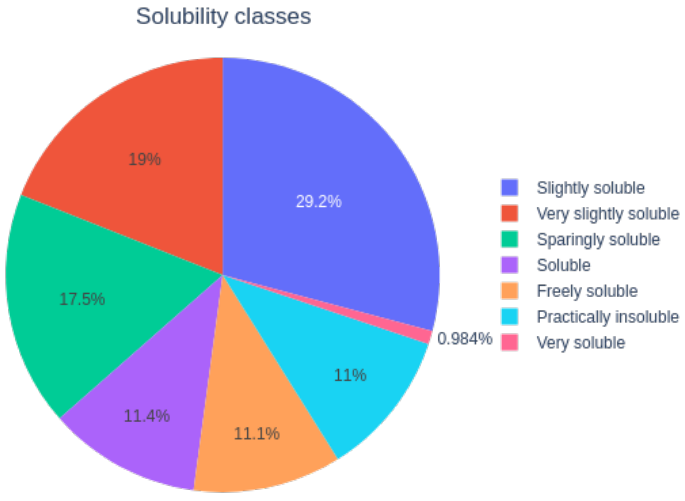

Figure 14. The distribution of USP solubility classes at 293.15 K