



ELSEVIER

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

Z-matrix template-based substitution approach for enumeration of 3D molecular structures



Wanutcha Lorpaiboon, Taweetham Limpanuparb*

Science Division, Mahidol University International College, Mahidol University, Salaya, Phutthamonthon, Nakhon Pathom 73170, Thailand

A B S T R A C T

The exhaustive enumeration of 3D chemical structures based on Z-matrix templates has recently been used in the quantum chemical investigation of constitutional isomers, diastereomers and rotamers. This simple yet powerful initial structure generation approach can apply beyond the investigation of compounds of identical formula by quantum chemical methods. This paper provides a comprehensive description of the overall concept followed by a practical tutorial to the approach.

- The four steps required for Z-matrix template-based substitution are template construction, generation of tuples for substitution sites, removal of duplicate tuples and substitution on the template.
- The generated tuples can be used to create chemical identifiers to query compound properties from chemical databases.
- All of these steps are demonstrated in this paper by common model compounds and are very straightforward for an undergraduate audience to reproduce. A comparison of the approach in this paper and other options is also discussed.

© 2021 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

A R T I C L E I N F O

Method name: Z-matrix template-based substitution*Keywords:* Chemical structures, Education, Molecular generator, Structure generator, Z-matrix*Article history:* Received 27 January 2021; Accepted 15 June 2021; Available online 17 June 2021

* Corresponding author.

E-mail address: taweetham.lim@mahidol.edu (T. Limpanuparb).

Specifications table

Subject Area:	Chemistry
More specific subject area:	Cheminformatics
Method name:	Z-matrix template-based substitution
Name and reference of original method:	N/A
Resource availability:	Source codes are available as supplementary information in this paper.

Introduction

Initial structures (Z-matrix or Cartesian coordinate) are important starting points for the *in silico* investigation of chemical species. Robust and exhaustive processes to generate chemical structures have been described previously in different contexts [1–9]. However, there are still no consistent one-size-fits-all standard for structural enumeration. Most of the existing methods are based on molecular graphs for representation of molecular structures and intrinsically lack crucial spatial information in terms of configuration and conformation of generated structures. Molecular graph-based methods may also suffer from resonance, unusual bond types or infeasible structures. There are ad hoc solutions to overcome these limitations [10] but, alternatively, a new method may be used instead or in combination with the existing methods.

Chemists, by their manual intuition and labour, have been able to identify Markush [33] or generic structures that represent a group of compounds with common chemical and/or physical properties. The Markush structures, represented as templates with defined substitution points, can be enumerated for the combinatorial investigation of compounds. Inspired by the use of this approach in dichlorodiphenyltrichloroethane (DDT) analogues [5] (Fig. 1), we successfully applied it to 26 classes of compounds to exhaustively generate over ten thousand 3D structures for further investigation by quantum chemical methods [11–15]. Based on these previous reports, this is our first paper devoted to the methodology of generic structure generation using Z-matrix template-based substitution. The concept and tutorial (template construction, generation of tuples for substitution sites, removal of duplicate tuples and substitution on the template) are described in the next section. Comparison of this approach to other existing methods and future work are discussed in the last section.

Z-matrix template-based substitution approach

Three dimensional molecular structures are often represented by a Z-matrix or Cartesian coordinates. The two representations can easily be interconverted by freely available software. We use Z-matrix for convenience as bond lengths, bond angles and torsional angles, which are natural quantities for chemists, can be explicitly laid out. (Cartesian coordinates would not be trivial to manipulate using this template-based substitution method.) As mentioned earlier, templates and substitution sites must be identified manually. For simplicity, in this article, we use hydrogen

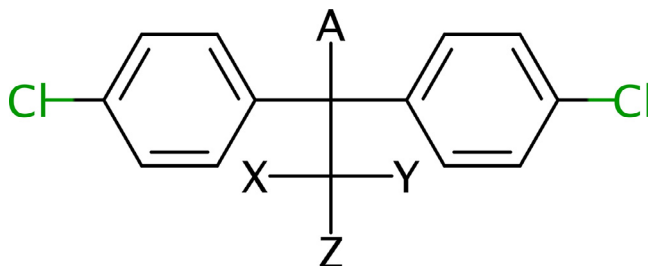


Fig. 1. Markush expression for DDT analogues taken for QSAR analysis: the analogues have structures represented by the template wherein A represents a member of a group consisting of OCH₃, H, F, and Br and X, Y, and Z represent members of a group consisting of CH₃, H, F, Cl, and Br.

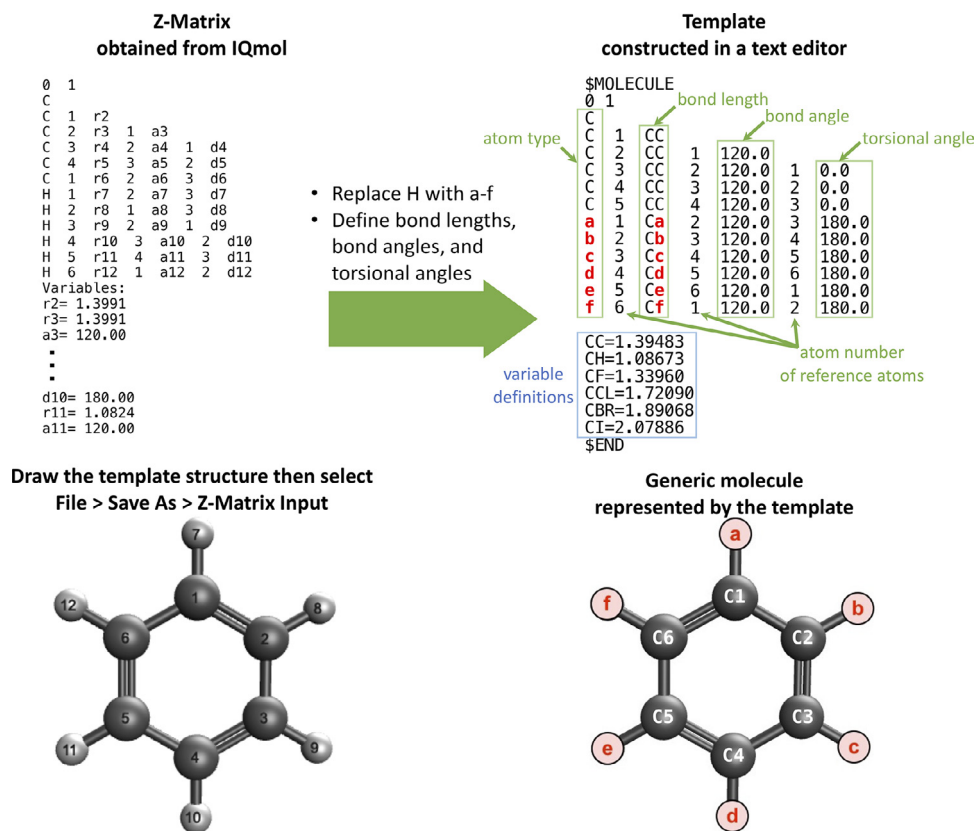


Fig. 2. Example of template for benzene in Q-Chem [21] input file format. Substitution positions (shown in red text) are labeled as a, b, c, d, e and f. H, F, Cl, BR, or I will take the place of each label after the process is complete. (Capital letters are used for atomic symbols in the Z-matrix to avoid complications in case-sensitive string replacement.) In the first line, the numbers 0 and 1 are the charge and multiplicity of the structure which are irrelevant to the discussion here but required by the file format. Each successive line provides the atom type and its relative position to three other atoms in the previous lines specified by bond length, bond angle and torsional angle.

and halogen atoms (F, Cl, Br and I) as substituents. The substitution approach can help us explore structures of different chemical compositions and structures of identical chemical formula (constitutional, configurational and conformational isomers) simultaneously.

In this paper, we demonstrate step-by-step how to generate an exhaustive collection of halo-substituted benzenes. Preferred toolkits on Windows platform are Notepad++ [16] for plain-text editing, Wolfram's Mathematica [17] for generation of tuples and duplicate elimination, GNU's sed [18] on Cygwin [19] for substitution into the template file, and IQmol [20] for visualization. These may be replaced by other software available on the user's platform.

Template construction

The first step of the procedure, construction of the template, can be achieved on any text editor. In this example, Notepad++ was used. The template molecule is defined as a Z-matrix with the substituting atoms in lowercase strings. (The use of lowercase letters is helpful for a case-sensitive replacement in later steps to avoid ambiguity with atomic symbols.) The example shown in Fig. 2 is a template for the generation of halo-substituted benzenes. The six halogen atom substitution sites

```
In[1]:= elements = {"H", "F", "Cl", "Br", "I"};
all = Tuples[elements, 6]
```

```
Out[ ]:= { {H, H, H, H, H, H}, {H, H, H, H, H, F}, {H, H, H, H, H, Cl},
  {H, H, H, H, H, Br}, {H, H, H, H, H, I}, {H, H, H, H, F, H}, {H, H, H, H, F, F},
  {H, H, H, H, F, Cl}, {H, H, H, H, F, Br}, {H, H, H, H, F, I}, {H, H, H, H, Cl, H},
  {H, H, H, H, Cl, F}, {H, H, H, H, Cl, Cl}, ... 15 600 ..., {I, I, I, I, Cl, Br},
  {I, I, I, I, Cl, I}, {I, I, I, I, Br, H}, {I, I, I, I, Br, F},
  {I, I, I, I, Br, Cl}, {I, I, I, I, Br, Br}, {I, I, I, I, Br, I}, {I, I, I, I, I, H},
  {I, I, I, I, I, F}, {I, I, I, I, I, Cl}, {I, I, I, I, I, Br}, {I, I, I, I, I, I} }
```

Fig. 3. Mathematica code used to generate tuples for substitution of benzene. A total of 5^6 or 15625 tuples were generated for halo-substituted benzenes. Duplicate entries are underlined using the same color.

are labelled as a, b, c, d, e and f. This template can be manually constructed by hand from scratch or by editing a Z-matrix from a molecular graphics software such as IQmol. For complicated molecules, it would be easier to draw or import the structure in such a program before manual manipulation. IQmol can help us visually identify the atom numbers corresponding to desired substitution sites by using Display > Atom Labels > Index. In this example, we see from IQmol that six hydrogen atoms are numbered from 7 to 12, and the letter H (for H atom) on the corresponding lines are changed to lowercase letters a to f accordingly. Bond lengths are defined as variables below the Z-matrix. (Bond and torsional angles can be defined as variables.) Therefore, a universal replacement of the letter a to a substituent atom (e.g. F) in step 4 will have two effects on the file; the type of atom (F) and bond length (CF) will be defined at the same time for the 7th atom in the Z-matrix.

Generation of tuples for substitution sites

The provided Wolfram Mathematica notebook (Fig. 3 and the Supplementary information) has a list named **elements** that contains the possible substituents. When generating halo-substituted benzenes, the possible substituents are H, F, Cl, Br and I atoms. Hence, **elements** = {"H", "F", "Cl", "Br", "I"}. **elements** can be modified when other substituting atoms or functional groups are studied. Tuples[**elements**, *k*] generates list **all** that contains all possible *k*-tuples from the elements in **elements**. The number of tuples generated in this step is equivalent to l^k where *l* is the length of **elements** and *k* is the size of each tuple. For benzene, possible substitutions for the slots a, b, c, d, e and f can be represented as a 6-tuple (*k* = 6). Therefore, the number of tuples generated from this step is $5^6 = 15625$.

Removal of duplicate tuples

The Mathematica notebook also contains operations to identify and remove duplicates and enantiomeric structures (Fig. 4). Each structure in the list of tuples **all** is rotated and flipped. If any of the resulting rotations or flips has been previously encountered, the corresponding tuple is discarded. Otherwise, the tuple is added to **finalist**. For example, the tuple {H, H, H, H, H, F} and {H, F, H, H, H, H} both refer to the same structure: fluorobenzene. Assuming the former comes earlier in **all**, it will be appended to **finalist**; the latter tuple will be discarded. For any number of substituents *n*, this method has a complexity of $O(n^2)$. The number of tuples in **finalist** can be verified mathematically [11–14].

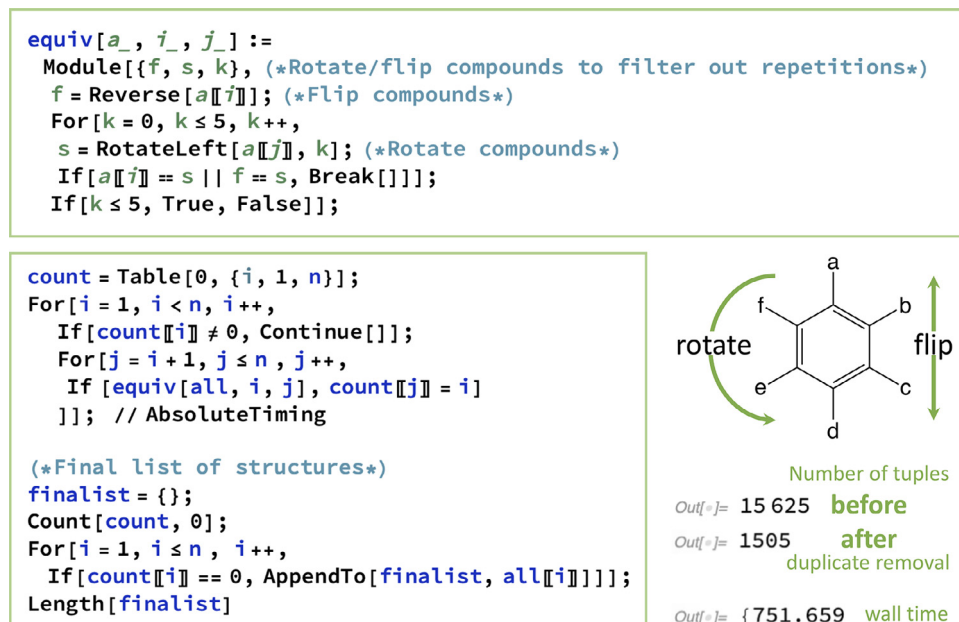


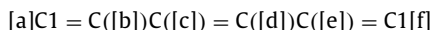
Fig. 4. Mathematica code used to eliminate identical and enantiomeric structures. The code used about 752 seconds of wall time and yielded 1505 unique structures. These results are consistent with combinatorial analysis of the halogenated benzene isomers [13].

Substitution on the template

Once the collection of tuples has been modified from the previous steps, GNU sed is used to replace the labels a, b, c, d, e and f in the template with appropriate atoms from the tuples in **finalist** (Fig. 5). The list of sed commands corresponding to all tuples are generated automatically from Mathematica. For each tuple, the substitution would result in a 3D chemical structure that can be visualized with any molecular graphics software and used as input files for further calculation.

Extension to chemical identifiers and database queries

Linear notations are simple descriptions of 3D structures and – more importantly – can be used as computer inputs. Many types of linear notations exist, namely the IUPAC international chemical identifier (InChI) and the simplified molecular-input line-entry system (SMILES). In this article, each tuple in **finalist** was converted into a SMILES string. As with the Z-matrix, a string template is constructed first. In the case of benzene, the template is



where the letters a to f can be substituted by elements in the tuple (Fig. 6). SMILES strings for each compound can be used to reference various properties of known compounds from databases.

Database queries provide quick look up of the compound properties and may help verify the uniqueness and the existence of the generated compounds. The National Institutes of Health maintains an open-access chemistry database called PubChem (<https://pubchem.ncbi.nlm.nih.gov>) [22, 23] that contains information such as chemical structures, identifiers, and properties of various small molecules and macromolecules. Each SMILES string can be submitted into the PubChem database to obtain a unique PubChem CID (Fig. 6). Duplicate compounds can be identified if any two SMILES strings return the same CID. This CID is also used to access compound properties in the database,

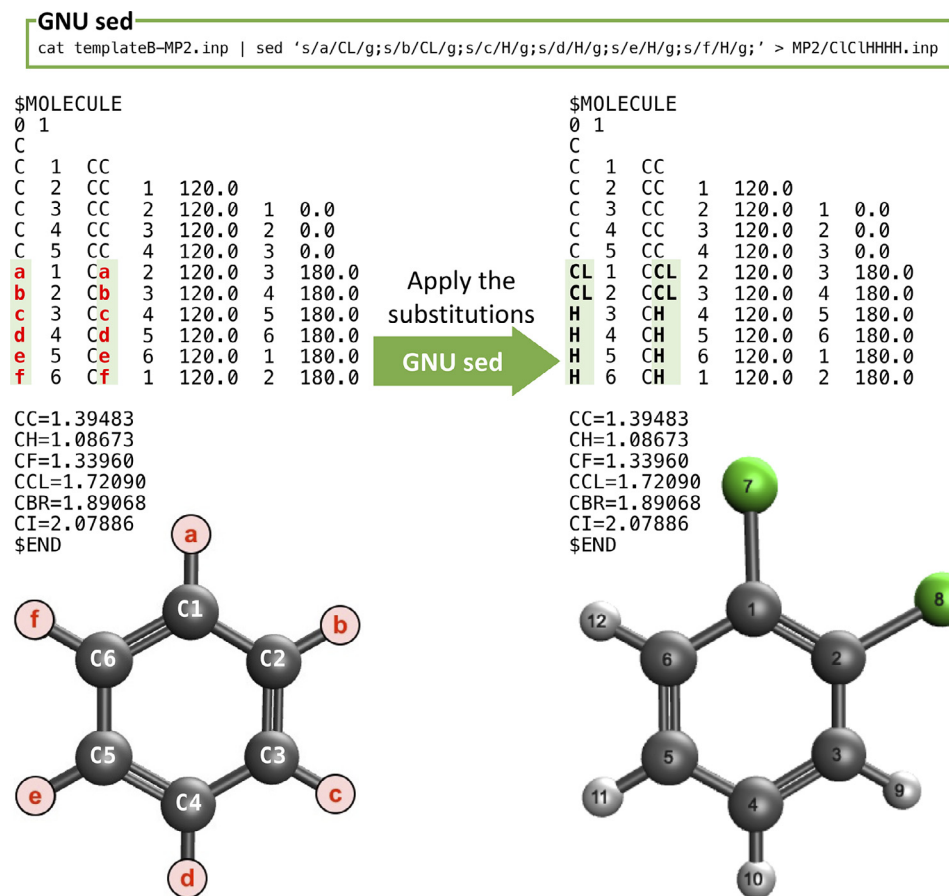


Fig. 5. Example of creating an input file from a template file by applying substitutions (yellow) using GNU sed. A 3D chemical structure of the file above is visualised in IQmol [20].

including molecular formula and weight, other linear notations such as InChI and IUPAC name, and exact mass. In addition, using the function `ChemicalData` in Mathematica [17], various properties such as boiling and melting points can be obtained. For instance, the relationship between molar masses and boiling points of 20 halogenated benzenes is shown as a scatter plot (Fig. 6). The ability to retrieve information from the database in bulk is an advantage in drug and smart materials discovery [24,25]. This also opens up possibilities, such as, analysis of chemical information by machine learning [26].

Discussion, conclusion and future work

This paper describes the concept of the Z-matrix template-based substitution approach and provides a simple practical tutorial to demonstrate the concept. This tutorial can be adapted to work on Mac, Linux and even Raspberry Pi by which Mathematica is free. There are also other alternative programs to implement the concept. For example, the code may be rewritten in other languages/platforms/frameworks such as Jupyter [27] and the results can be visualized in Avogadro [28] or GaussView [34].

The approach has been used to study many classes of compounds from a simple and small class of molecules such as ethane and ethene to larger and more complex structures such as dioxin-like

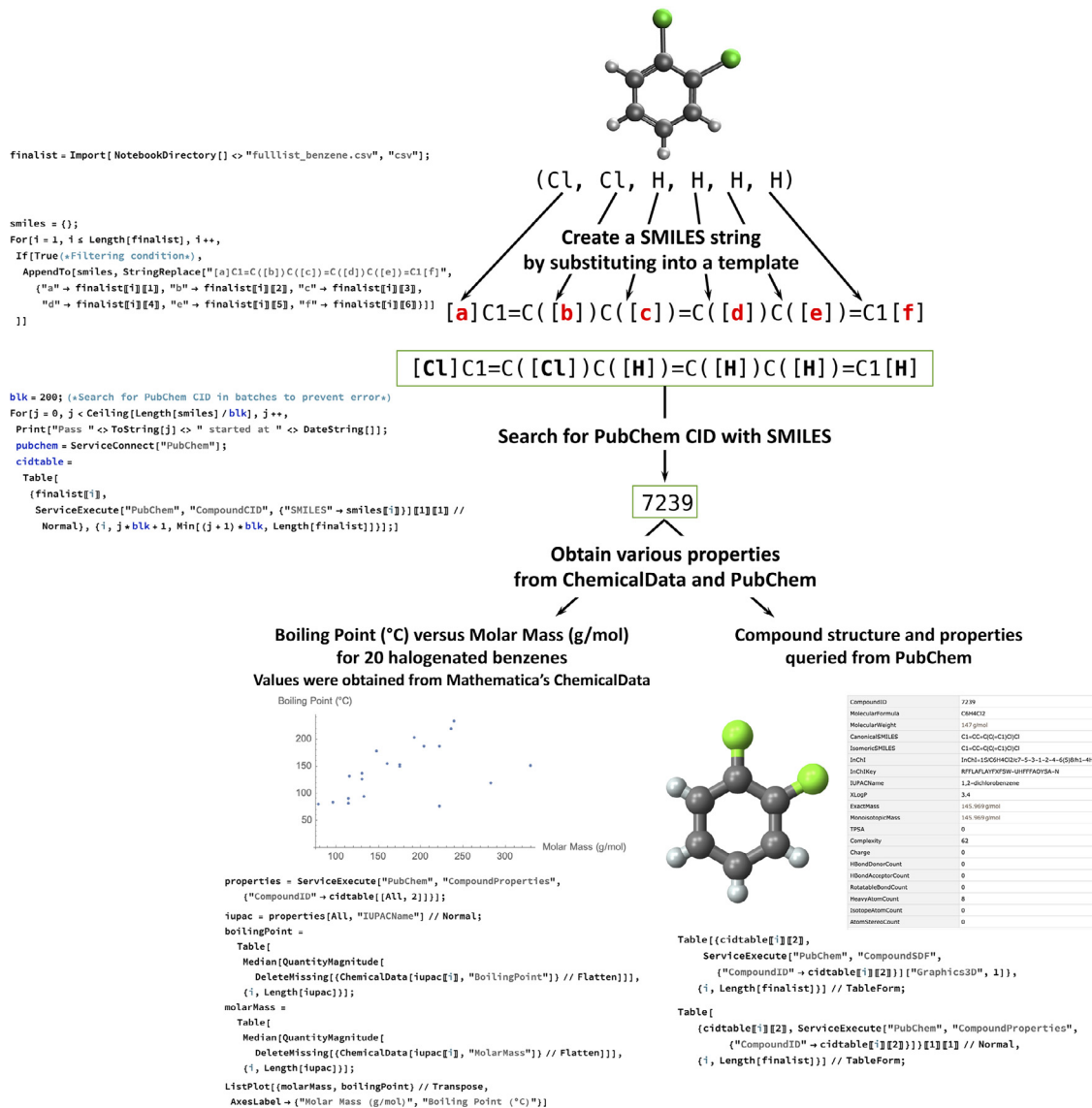
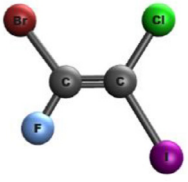
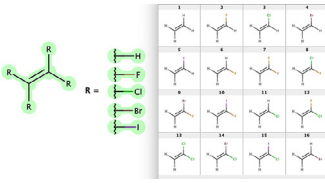
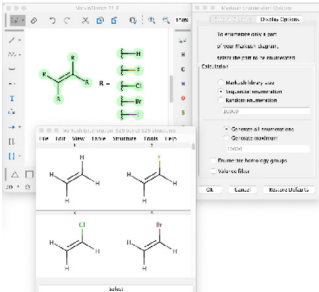
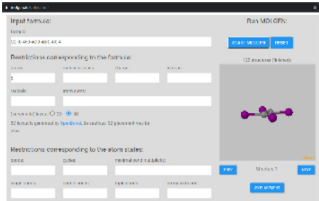


Fig. 6. Translation of each tuple to a SMILES string and PubChem CID. The PubChem CID can be used to look up compound information in the PubChem database.

Table 1
Comparison of the discussed approach and available software in the market.

Software	Screenshot of results	Description
Z-matrix template-based substitution approach		A total of 175 unique structures were correctly generated [11].
ChemDraw 20.1 [29]		ChemDraw initially generated 625 structures, followed by removal of duplicates and finally displayed (generally only up to 200 structures) 181 structures. There were still duplicate entries in the generated structures. It is not clear why a majority of the duplicates were removed but only 6 pairs remained.
MarvinSketch 21.8 [30]		MarvinSketch produced 625 structures because the program did not remove any duplicates.
MOLGEN 5.0 (online) [31]		MOLGEN quickly returned 120 structures for 2D option and took seconds for the 3D option. This result of 120 structures is intentional by design. The program generates connectivity isomers, and, as a result, it does not account for stereoisomers.

compounds [11–15]. The substitution scheme in the template can lead to different compounds of the same class, constitutional isomers, configurational isomers and conformational isomers. However, given limitation of other existing off-the-shelf software, a relatively simple class of halogenated ethenes is selected as a model example for comparison as shown in Table 1. The three programs yielded 6 extra, 450 extra or 55 missing entries compared to our expected outcome of 175 entries. (See supplementary information for more details. We also attempted on some codes/workflows on popular KNIME [32] but did not manage to produce any results for this example case.) Therefore, this example of halogenated ethenes raises a very important concern that are typical of molecular databases and libraries. The integrity of data can be improved if initial structures from upstream are generated correctly.

Our approach presented here for educational purposes is straightforward and may be used in classrooms for training. The codes are available in the Supplementary information and expected to be modified for specialized uses beyond tutorial purposes. Computer coding/scripting is inevitably an integral part of this tutorial. Fortunately, by expressing our ideas in Mathematica, the number of lines

of code can be fairly at minimum. The natural next step after this tutorial is filtering for molecular structures in which desired conditions are fulfilled. For example, in our previous work, we labelled structures as *E* or *trans*, *Z* or *cis*, geminal, *gauche*, *anti*, *ortho*, *meta*, *para*. Custom-made codes/scripts are freely available in the references [11–15] and may be modified to fit the needs of users.

Two limitations that can be addressed in the future are the use of non-single atom substituents and the efficiency of the duplicate removal step. The first limitation will generally require a consideration of the torsional angles of the substituent groups with respect to the template. The second limitation may be overcome by using a different computer language, sacrificing the simplicity and brevity of Mathematica codes. An alternative approach is to represent each structure as a SMILES string and filter duplicates using the PubChem database. For example, searching C1C1=CC=CC(F)=C1 and FC1=CC(C1)=CC=C1 returns the same compound: 1-chloro-3-fluorobenzene. Hence, one of the structures is a duplicate and can be removed. Although this approach may only work for compounds that are identified, PubChem also allows quick access to the properties of the compounds which is particularly useful when enumerating structures for drug discovery and smart materials.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

T.L. is thankful for the opportunity to present this work at the Development and Promotion of Science and Technology Talents Project (DPST) Conference on Science and Technology on 4 July 2020. We are grateful to Kritdin Chinsukserm and Sopianant Datta for analysis, ideas and suggestions in the early phase of this project. We thank Markus Meringer for his comments on this manuscript.

Author contributions

Conceptualization, W.L. and T.L.; data curation, W.L. and T.L.; formal analysis, W.L.; funding acquisition, T.L.; investigation, T.L.; methodology, W.L. and T.L.; project administration, T.L.; resources, T.L.; software, T.L.; supervision, T.L.; validation, W.L. and T.L.; visualization, W.L.; writing—original draft preparation, W.L. and T.L.; writing—review and editing, W.L. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by Mid-Career Researcher Development grant (NRCT5-RSA63015-22) jointly funded by the National Research Council of Thailand (NRCT) and Mahidol University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Data and materials for the examples are available as additional materials.

Supplementary materials

Supplementary data associated with this article can be found, in the online version, at doi:[10.1016/j.mex.2021.101416](https://doi.org/10.1016/j.mex.2021.101416).

References

- [1] M. Randić, C.L. Wilkins, Graph-theoretical analysis of molecular properties. Isomeric variations in nonanes, *Int. J. Quantum Chem.* 18 (1980) 1005–1027.
- [2] M. Randić, Graphical enumeration of conformations of chains, *Int. J. Quantum Chem.* 18 (1980) 187–197.
- [3] J.-F. Truchon, C.I. Bayly, GLARE: a new approach for filtering large reagent lists in combinatorial library design using product properties, *J. Chem. Inf. Model.* 46 (2006) 1536–1548.
- [4] J.E. Peironcely, M. Rojas-Chertó, D. Fichera, T. Reijmers, L. Coulier, J.-L. Faulon, T. Hankemeier, OMG: Open Molecule Generator, *J. Cheminformatics* 4 (2012) 21.
- [5] V. Saini, A. Kumar, QSAR analyses of DDT analogues and their in silico validation using molecular docking study against voltage-gated sodium channel of *Anopheles funestus*, *SAR QSAR Environ. Res.* 25 (2014) 777–790.
- [6] D. Thiagarajan, D.P. Mehta, Faster algorithms for isomer network generation, *J. Chem. Inf. Model.* 56 (2016) 2310–2319.
- [7] M. Koch, T. Duigou, P. Carbonell, J.-L. Faulon, Molecular structures enumeration and virtual screening in the chemical space with RetroPath2.0, *J. Cheminformatics* 9 (2017) 64.
- [8] A. Voicu, N. Duteanu, M. Voicu, D. Vlad, V. Dumitrascu, The rcdk and cluster R packages applied to drug candidate selection, *J. Cheminformatics* 12 (2020) 3.
- [9] F.I. Saldívar-González, C.S. Huerta-García, J.L. Medina-Franco, Chemoinformatics-based enumeration of chemical libraries: a tutorial, *J. Cheminformatics* 12 (2020) 64.
- [10] J.-M. Gally, S. Bourg, Q.-T. Do, S. Ací-Sèche, P. Bonnet, VSPrep: A General KNIME workflow for the preparation of molecules for virtual screening, *Mol. Inform.* 36 (2017) 1700023.
- [11] K. Chinsukserm, W. Lorpaiboon, P. Teeraniramit, T. Limpanuparb, Geometric and energetic data from ab initio calculations of haloethene, haloimine, halomethylenephosphine, haloiminophosphine, halodiazene, halodiphosphene and halocyclopropane, *Data Brief* 27 (2019) 104738.
- [12] S. Datta, T. Limpanuparb, Quantum chemical investigation of polychlorinated dibenzodioxins, dibenzofurans and biphenyls: relative stability and planarity analysis, *Molecules* 25 (2020) 5697.
- [13] S. Datta, T. Limpanuparb, Geometric and energetic data from quantum chemical calculations of halobenzenes and xylenes, *Data Brief* (2020) 105386.
- [14] T. Limpanuparb, S. Datta, K. Chinsukserm, P. Teeraniramit, In silico geometric and energetic data of all possible simple rotamers made of non-metal elements, *Data Brief* (2020) 105442.
- [15] S. Datta, T. Limpanuparb, Steric effects vs. electron delocalization: a new look into the stability of diastereomers, conformers and constitutional isomers, *RSC Adv* 11 (2021) 20691–20700 submitted.
- [16] D. Ho, Notepad++, 2020.
- [17] Wolfram Research Inc. Mathematica, Wolfram Research Inc., Champaign, Illinois, 2021.
- [18] sed, a stream editor.
- [19] Cygnus Solutions, Cygwin, 2020.
- [20] A. Gilbert, IQmol, 2019.
- [21] Y. Shao, Z. Gan, E. Epifanovsky, A.T. Gilbert, M. Wormit, J. Kussmann, et al., Advances in molecular quantum chemistry contained in the Q-Chem 4 program package, *Mol. Phys.* 113 (2015) 184–215.
- [22] 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 in 2021: new data content and improved web interfaces, *Nucleic Acids Res* 49 (2020) D1388–D1395.
- [23] 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 2019 update: improved access to chemical data, *Nucleic Acids Res* 47 (2018) D1102–D1109.
- [24] A. Daina, M.-C. Blatter, V. Baillie Gerritsen, P.M. Palagi, D. Marek, I. Xenarios, T. Schwede, O. Michielin, V. Zoete, Drug design workshop: a web-based educational tool to introduce computer-aided drug design to the general public, *J. Chem. Educ.* 94 (2017) 335–344.
- [25] V.V. Acuna, R.M. Hopper, R.J. Yoder, Computer-aided drug design for the organic chemistry laboratory using accessible molecular modeling tools, *J. Chem. Educ.* 97 (2020) 760–763.
- [26] L. Joss, E.A. Müller, Machine learning for fluid property correlations: classroom examples with MATLAB, *J. Chem. Educ.* 96 (2019) 697–703.
- [27] T. Kluyver, B. Ragan-Kelley, F. Pérez, B.E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J.B. Hamrick, J. Grout, S. Corlay, Jupyter Notebooks—a publishing format for reproducible computational workflows, *ELPUB* (2016) 87–90.
- [28] M.D. Hanwell, D.E. Curtis, D.C. Lonie, T. Vandermeersch, E. Zurek, G.R. Hutchison, Avogadro: an advanced semantic chemical editor, visualization, and analysis platform, *J. Cheminformatics* 4 (2012) 17.
- [29] PerkinElmer Informatics, ChemDraw 20.1.0.110, 2021.
- [30] ChemAxon, MarvinSketch (version 21.8.0, calculation module developed by, ChemAxon) (2021).
- [31] R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker, A. Wassermann, Chapter 6 - MOLGEN 5.0, a molecular structure generator, in: S.C. Basak, G. Restrepo, J.L. Villaveces (Eds.), *Advances in Mathematical Chemistry and Applications*, Bentham Science Publishers, 2015, pp. 113–138.
- [32] M.R. Berthold, N. Cebron, F. Dill, T. Gabriel, KNIME - the konstanz information miner: version 2.0 and beyond, *SIGKDD Explor. Newsl.* 11 (2009) 26–31.
- [33] E.H. Valance, Understanding the Markush claim in chemical patents, *J Chem Doc* 1 (1961) 87–92.
- [34] R. Dennington, T.A. Keith, J.M. Millam, GaussView, Semichem Inc, Shawnee Mission KS, 2021.