Journal of Cheminformatics

SOFTWARE Open Access

ChemEngine: harvesting 3D chemical structures of supplementary data from PDF files

Muthukumarasamy Karthikeyan[1*] and Renu Vyas[2]

Abstract

Digital access to chemical journals resulted in a vast array of molecular information that is now available in the supplementary material files in PDF format. However, extracting this molecular information, generally from a PDF document format is a daunting task. Here we present an approach to harvest 3D molecular data from the supporting information of scientific research articles that are normally available from publisher's resources. In order to demonstrate the feasibility of extracting truly computable molecules from PDF file formats in a fast and efficient manner, we have developed a Java based application, namely ChemEngine. This program recognizes textual patterns from the supplementary data and generates standard molecular structure data (bond matrix, atomic coordinates) that can be subjected to a multitude of computational processes automatically. The methodology has been demonstrated via several case studies on different formats of coordinates data stored in supplementary information files, wherein ChemEngine selectively harvested the atomic coordinates and interpreted them as molecules with high accuracy. The reusability of extracted molecular coordinate data was demonstrated by computing Single Point Energies that were in close agreement with the original computed data provided with the articles. It is envisaged that the methodology will enable large scale conversion of molecular information from supplementary files available in the PDF format into a collection of ready- to- compute molecular data to create an automated workflow for advanced computational processes. Software along with source codes and instructions available at https://sourceforge.net/projects/chemengine/files/?source=navbar.

Keywords: Chemoinformatics, Supplementary data, Organic reaction modeling, Density functional theory, Text mining, Data mining

Background

Harvesting chemical data from the web is a challenging task requiring several convoluted steps. When chemical structures are stored in truly computable format with atoms and bond matrices (vector format-Cartesian co-ordinates), they can be processed electronically for computational and informatics purposes. However while transforming/storing the files in PDF (Printable/Portable Document/Data Format) that are usually used for the convenience of printing and reading, the valuable and re-usable molecular data is totally lost and buried in scientific literature as documents and seldom used for further computational studies. In earlier days, the hand-drawn molecules in ORTEP diagram formats were published while discussing the 3D conformation of molecules in the research articles. Generation of 3D structures from these molecular images in raster format was extremely difficult. Recently, some efforts have been made to transform computer generated and hand-drawn chemical images from journal articles and patent documents into truly computable molecules for inventory and database applications. Other similar endeavors include transforming either the textual chemical names (common, systematic, corporate identifiers for example CAS Registry number) or the computer generated names into corresponding molecular structures with moderate success. Although the name to chemical structure conversion programs are now routinely being used for harvesting chemical data from documents yet they have been insufficient in generating

*Correspondence: m.karthikeyan@ncl.res.in
[1] Present Address: Chemical Engineering and Process Development (CEPD), CSIR-National Chemical Laboratory, Pashan Road, Pune, Maharastra 411008, India
Full list of author information is available at the end of the article

the accurate and truly computable and re-usable molecular data. The supporting information related to computational methods based research articles, describing the transition states of organic reactions is now available from journal publishers' websites containing description of computations performed with tables of results, molecular images in 3D conformations along with 3D molecular co-ordinates in a PDF format. This combined data in a single file complicates the harvesting process and development of pattern recognition techniques for selectively excluding the non-atomic co-ordinate information from the pool of large collection of textual data presented as supporting material. Since there are no defined rules and guidelines for submitting molecular data in a supporting document associated with research publications, the authors are free to choose their favorite methods of representing molecular data such as chemical structures and corresponding atomic co-ordinates in the supplementary data file. This freedom of choosing data formats necessitates the development of several pattern recognition templates in the form of regular expressions to handle diverse formats (co-ordinates separated by space, comma, tab etc.) and maintain the order in which the XYZ co-ordinates and atom information is presented by the authors. This study therefore highlights the need for development of standards required for submitting the supporting materials with molecular data in a consistent, truly computable and re-usable format to journals publishing computational research. A specific set of guidelines defined by the publishers to submit molecular data even in a PDF format, would accelerate the automatic processing and recognition of chemical data for further computational studies related to reaction modeling [1–3], drug-discovery [4–7] and molecular inventory management [8, 9]. Several standard molecular representations in ASCII format which are easily readable by molecular modeling and chemoinformatics software packages are available. Supporting materials are deposited in PDF format for the convenience of storage, easy manageability and electronic dissemination. The commercial software packages applied for computational chemistry applications employ their own legacy file formats for handling molecular data, the technical details of which are not usually published. From the researchers' point of view, the published data in re-usable formats would save efforts and time to understand the molecular data better and use it for practicing to carry out further advanced studies in different problem solving environments that require 3D conformation of molecules. Exchange of chemical data between multiple softwares without loss of information is a critical requirement in computational chemistry and chemoinformatics applications. Thus there is a need for the development of tools that can bridge the gap in molecular data translation automatically and accurately from PDF format to truly computable, re-usable format without manual intervention.

In this context, it is pertinent to mention the efforts by Rzepa and Peter Murray-Rust for developing tools to parse chemically relevant thesis and other published articles for harvesting analytical data [10, 11]. Special emphasis was laid on the use of Information Technology (IT) techniques for free re-distribution of electronic chemical data, for instance, storing actual supplementary information in structured XML/CML documents for universal applicability and dissemination of the valuable experimental/computed data thus advancing "data led science" as is the case in biology. The blue obelisk informal group initiative [12], encourages the use of open source data, open standards, shared algorithms and tools for performing chemoinformatics tasks. It has led to the development of valuable tools such as JChemPaint [13], CDK [14] and chemical information systems [15]. Similar efforts have been made by the Cambridge Crystallographic Data Center (CCDC) group that provides easily downloadable crystal structures of organic molecules that are pliant with a number of software solutions for drug discovery [16]. In a recent article, the importance of curation of large chemogenomics data set for building better predictive model for life sciences has been emphasized [17]. During the preparation of this manuscript, a timely research article by Rzepa's group on granularity model for extracting molecular information appeared [18] that stresses on the need for periodic and automatic curation of data from supplementary information in research articles. The present work is geared towards partial fulfillment of this need for "futuristic research data management".

Conventionally, chemical names (common, systematic), Chemical Abstract Registry numbers are extracted from the web-pages and transformed into corresponding molecular structures using name-to-structure conversion tools [19], name to structure relational database look-up methods [20], large scale key-value pair list [21], distributed relational database search [22] etc. We have previously employed distributed systems to harvest chemical data using Google API (ChemXtreme) from the web pages [23]. Transforming the raster images into vector graphics followed by identification of relevant pixel information associated with atoms and bonds of a molecule is a cumbersome job [24]. Tools have also been developed to harvest molecular data from images using web camera, scanned images wherein the raster graphics data was transformed into vector graphics to eventually retrieve the atoms and bonds information for the generation of truly computable and re-usable chemical structures such as ChemRobot [25], OSRA [26], ChemReader [27], CLiDE [28], but only limited success has been achieved. A foolproof method with complete reproducibility of computable molecules from images is still a distant dream as the existing methodologies and tools do not provide accurate molecule data after processing. Therefore it is essential to develop efficient tools that can extract molecules from rich sources such as supplementary information files deposited at the journal site. Although spectral, molecular and analytical data have been harvested in the past but extracting molecules directly from author supplied atomic coordinates provided in supplementary materials as PDF format is not known. Accordingly, in the present work, we have developed an application, ChemEngine that reads all the files stored in the PDF format to extract molecular coordinates and generate computable molecular structures. To demonstrate the efficiency of the program, supporting material data files of three different molecular representations in terms of delimiters in the co-ordinate data were selected and the data was successfully parsed using ChemEngine to extract molecular data. It is to be noted here that the first two files from ACS publications did not require permission for data harvesting, while in the third case (RSC Advances), an article published under the CC-BY license was selected. It is also observed that the bulk processing of articles or supporting materials from publishers' site automatically is usually prohibited due to copyright and article access policy.

Generally every software program dealing with computational chemistry, provides an export format for the computed data either as a plain text or delimited text that can be analyzed, visualized, plotted via common tools like Microsoft excel or other molecular viewers that accept molecules as plain text in simple.xyz formats. However, supporting materials of molecular data files also include brief description of molecules, computed data, plots, page numbers, document information, manuscript bibliographic details etc. as a single document in PDF format that makes harvesting the molecular data extremely difficult as these have to be selectively excluded while parsing the file. In the Fig. 1, only the enclosed text in the rectangular box is correctly recognized using patterns by ChemEngine, the rest of the unstructured text is ignored. Given an input file in PDF format, the program yields three different files in GJF format, text file containing computed bond matrix and all molecules in SDF format. The contents of the non molecular data file can also be utilized by further subjecting it to standard text mining methodologies [29, 30] for retrieving molecule names or other information such as list of basis sets employed in the specific computational work.

Implementation

The ChemEngine application was deployed on a computer with Intel Xeon(R) CPU E5-2603 (2 Processors), 16 GB RAM, 1 TB hard disk running 64 bit operating system on Windows Server 2008 Enterprise. The computational steps employed in the Java based ChemEngine application are highlighted in the flowchart (Fig. 2).

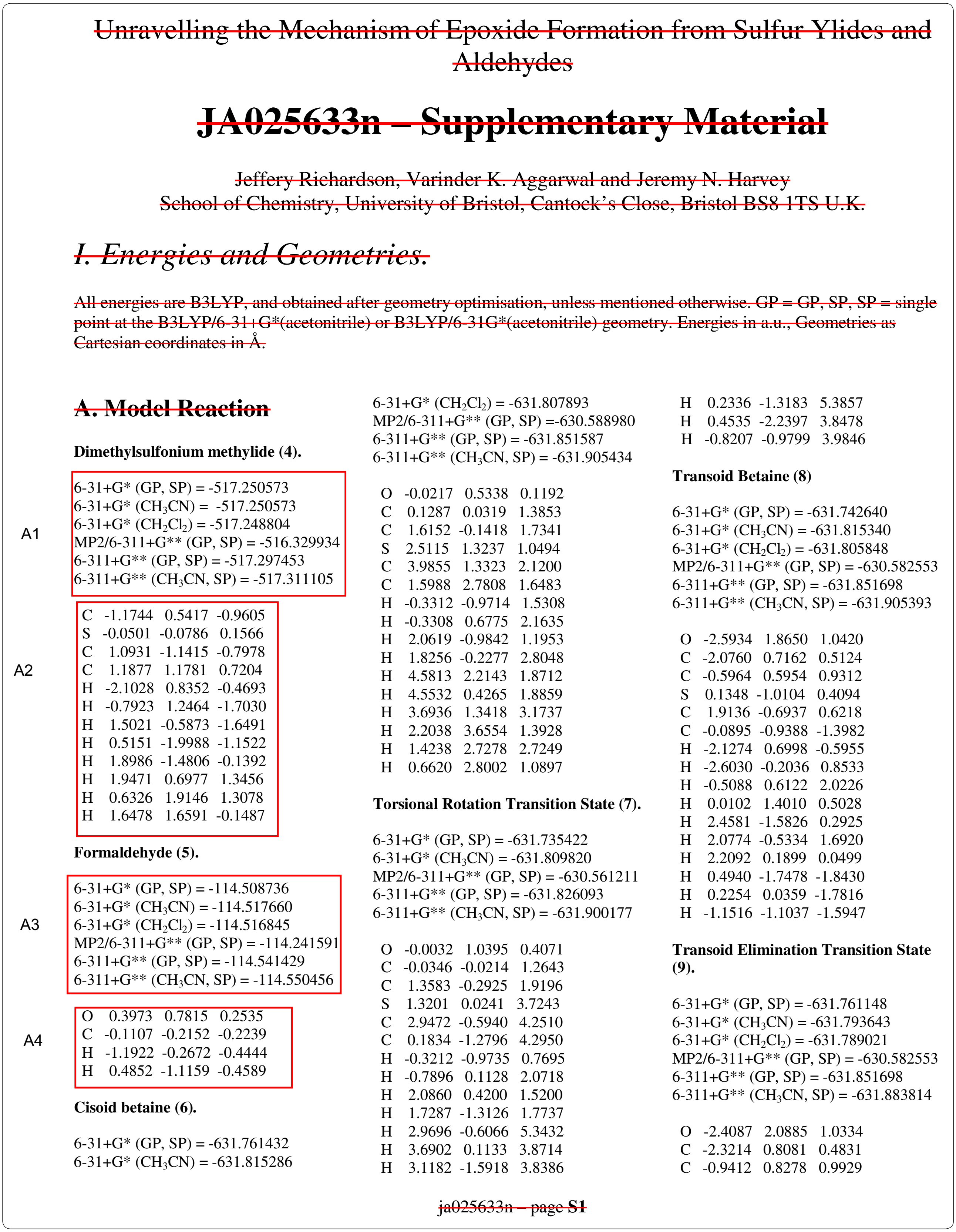
ChemEngine is updated with default option to accept PDF file containing molecular coordinates. Internally the program recognizes the textual and non-textual data and using a default pattern recognition method to separate the 3D coordinates from the non-molecular text for the identification of atomic co-ordinates and atom information. The pseudo code with generic regular expression for harvesting atomic coordinate data from the input file is shown below.

Pseudo code

```
(Co-ordinate Text).matches ("Regular Expression Pattern with Delimiter Definition");
    For Example: Delimiter: Comma
    String_Data.matches("^[A-Za-z0-9]{1,2}\\,[0]{0,1}[\\,]{0,1}-{0,1}.{1,2}[0-9]{1,10}\\,-{0,1}.{1,2}[0-9]{1,10}.{1,}")
    Delimiter: Space
    String_Data.matches("^[A-Za-z0-9]{1,2}\\s+[0]{0,1}[\\s +]{0,1}-{0,1}.{1,2}[0-9]{0,10}\\s+-{0,1}.{1,2}[0-9]{0,10}.{1,}")
```

The details of the derivation of the regular expression patterns from the coordinate data format can be viewed in Fig. 3. All the X, Y, Z coordinates were encoded by a general pattern sequence consisting of 2 characters, followed by a space, an addition or subtraction symbol, a number, decimal and eight digits succeeding the decimal.

Once the coordinate file is created, the bond matrix is computed to generate the atomic connectivity information for reconstructing the original molecules reported in the supplementary material of the research article. Important parameters such as bond angles, bond lengths and dihedral angles are verified and checked for consistency in the recreated molecule and then saved in the original file format, for instance gjf [31]. The coordinate data and bond matrix information is used to create molecules in standard interoperability formats such as .sdf or .mol as ready to compute molecules for the convenience of the user. This process avoids unnecessary generation of molecular data and laborious recomputation of already published work. The molecules can be subjected to further simulations such as descriptor calculation, energy profile, docking etc. The java based ChemEngine program is made available freely for non commercial purposes through the sourceforge site for evaluation and testing.



~~Unravelling the Mechanism of Epoxide Formation from Sulfur Ylides and Aldehydes~~

~~JA025633n – Supplementary Material~~

~~Jeffery Richardson, Varinder K. Aggarwal and Jeremy N. Harvey~~
~~School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS U.K.~~

~~*I. Energies and Geometries.*~~

~~All energies are B3LYP, and obtained after geometry optimisation, unless mentioned otherwise. GP = GP, SP, SP = single point at the B3LYP/6-31+G*(acetonitrile) or B3LYP/6-31G*(acetonitrile) geometry. Energies in a.u., Geometries as Cartesian coordinates in Å.~~

~~A. Model Reaction~~

Dimethylsulfonium methylide (4).

A1

6-31+G* (GP, SP) = -517.250573
6-31+G* (CH_3CN) = -517.250573
6-31+G* (CH_2Cl_2) = -517.248804
MP2/6-311+G** (GP, SP) = -516.329934
6-311+G** (GP, SP) = -517.297453
6-311+G** (CH_3CN, SP) = -517.311105

A2

C	-1.1744	0.5417	-0.9605
S	-0.0501	-0.0786	0.1566
C	1.0931	-1.1415	-0.7978
C	1.1877	1.1781	0.7204
H	-2.1028	0.8352	-0.4693
H	-0.7923	1.2464	-1.7030
H	1.5021	-0.5873	-1.6491
H	0.5151	-1.9988	-1.1522
H	1.8986	-1.4806	-0.1392
H	1.9471	0.6977	1.3456
H	0.6326	1.9146	1.3078
H	1.6478	1.6591	-0.1487

Formaldehyde (5).

A3

6-31+G* (GP, SP) = -114.508736
6-31+G* (CH_3CN) = -114.517660
6-31+G* (CH_2Cl_2) = -114.516845
MP2/6-311+G** (GP, SP) = -114.241591
6-311+G** (GP, SP) = -114.541429
6-311+G** (CH_3CN, SP) = -114.550456

A4

O	0.3973	0.7815	0.2535
C	-0.1107	-0.2152	-0.2239
H	-1.1922	-0.2672	-0.4444
H	0.4852	-1.1159	-0.4589

Cisoid betaine (6).

6-31+G* (GP, SP) = -631.761432
6-31+G* (CH_3CN) = -631.815286
6-31+G* (CH_2Cl_2) = -631.807893
MP2/6-311+G** (GP, SP) =-630.588980
6-311+G** (GP, SP) = -631.851587
6-311+G** (CH_3CN, SP) = -631.905434

O	-0.0217	0.5338	0.1192
C	0.1287	0.0319	1.3853
C	1.6152	-0.1418	1.7341
S	2.5115	1.3237	1.0494
C	3.9855	1.3323	2.1200
C	1.5988	2.7808	1.6483
H	-0.3312	-0.9714	1.5308
H	-0.3308	0.6775	2.1635
H	2.0619	-0.9842	1.1953
H	1.8256	-0.2277	2.8048
H	4.5813	2.2143	1.8712
H	4.5532	0.4265	1.8859
H	3.6936	1.3418	3.1737
H	2.2038	3.6554	1.3928
H	1.4238	2.7278	2.7249
H	0.6620	2.8002	1.0897

Torsional Rotation Transition State (7).

6-31+G* (GP, SP) = -631.735422
6-31+G* (CH_3CN) = -631.809820
MP2/6-311+G** (GP, SP) = -630.561211
6-311+G** (GP, SP) = -631.826093
6-311+G** (CH_3CN, SP) = -631.900177

O	-0.0032	1.0395	0.4071
C	-0.0346	-0.0214	1.2643
C	1.3583	-0.2925	1.9196
S	1.3201	0.0241	3.7243
C	2.9472	-0.5940	4.2510
C	0.1834	-1.2796	4.2950
H	-0.3212	-0.9735	0.7695
H	-0.7896	0.1128	2.0718
H	2.0860	0.4200	1.5200
H	1.7287	-1.3126	1.7737
H	2.9696	-0.6066	5.3432
H	3.6902	0.1133	3.8714
H	3.1182	-1.5918	3.8386
H	0.2336	-1.3183	5.3857
H	0.4535	-2.2397	3.8478
H	-0.8207	-0.9799	3.9846

Transoid Betaine (8)

6-31+G* (GP, SP) = -631.742640
6-31+G* (CH_3CN) = -631.815340
6-31+G* (CH_2Cl_2) = -631.805848
MP2/6-311+G** (GP, SP) = -630.582553
6-311+G** (GP, SP) = -631.851698
6-311+G** (CH_3CN, SP) = -631.905393

O	-2.5934	1.8650	1.0420
C	-2.0760	0.7162	0.5124
C	-0.5964	0.5954	0.9312
S	0.1348	-1.0104	0.4094
C	1.9136	-0.6937	0.6218
C	-0.0895	-0.9388	-1.3982
H	-2.1274	0.6998	-0.5955
H	-2.6030	-0.2036	0.8533
H	-0.5088	0.6122	2.0226
H	0.0102	1.4010	0.5028
H	2.4581	-1.5826	0.2925
H	2.0774	-0.5334	1.6920
H	2.2092	0.1899	0.0499
H	0.4940	-1.7478	-1.8430
H	0.2254	0.0359	-1.7816
H	-1.1516	-1.1037	-1.5947

Transoid Elimination Transition State (9).

6-31+G* (GP, SP) = -631.761148
6-31+G* (CH_3CN) = -631.793643
6-31+G* (CH_2Cl_2) = -631.789021
MP2/6-311+G** (GP, SP) = -630.582553
6-311+G** (GP, SP) = -631.851698
6-311+G** (CH_3CN, SP) = -631.883814

O	-2.4087	2.0885	1.0334
C	-2.3214	0.8081	0.4831
C	-0.9412	0.8278	0.9929

~~ja025633n – page S1~~

(See figure on previous page.)
Fig. 1 Supplementary data of a journal article (case study I) depicting the computed molecular data format, the contents in the *highlighted text* are required for the re-computation of data. A1, A2, B1, B2 refer to text patterns in the specific document. The *crossed out text in red color* is ignored while generating the coordinate file by ChemEngine version 1.0

A user friendly GUI has been developed for easy processing of PDF files, navigation and 3D structure generation (Fig. 4). The ChemEngine main screen displays the input file browser, the output display text area and a text box for specify regular expression. The GUI is enabled with a 3D molecule viewer (JMOL) for browsing the molecules in a pop up window. The program can also be used in the command line mode for automatic generation of SDF for the given PDF files. The user can either browse a locally stored PDF file or upload a text file for generating coordinates. Depending upon how the coordinates data is deposited in a supplementary file, there is

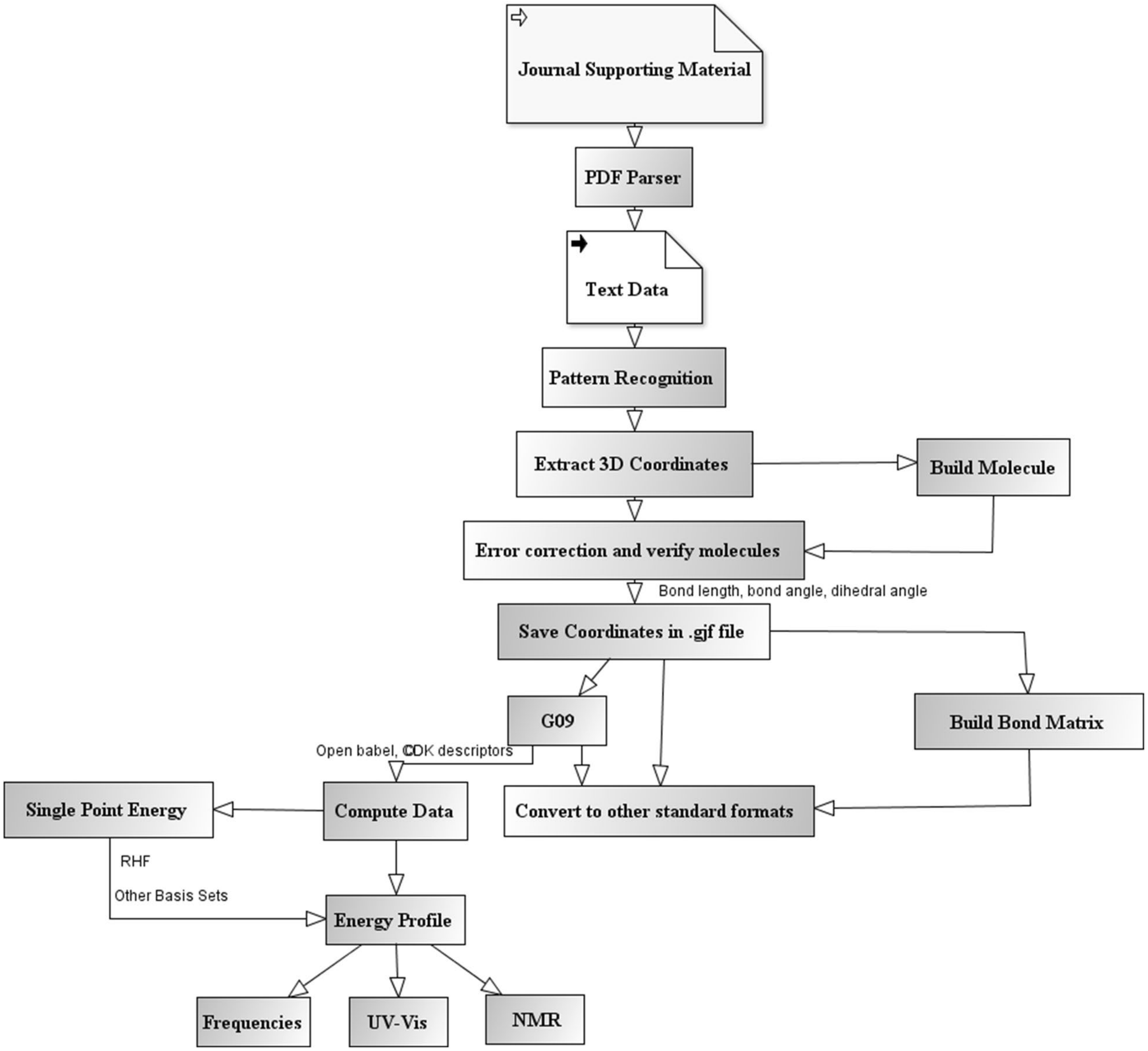
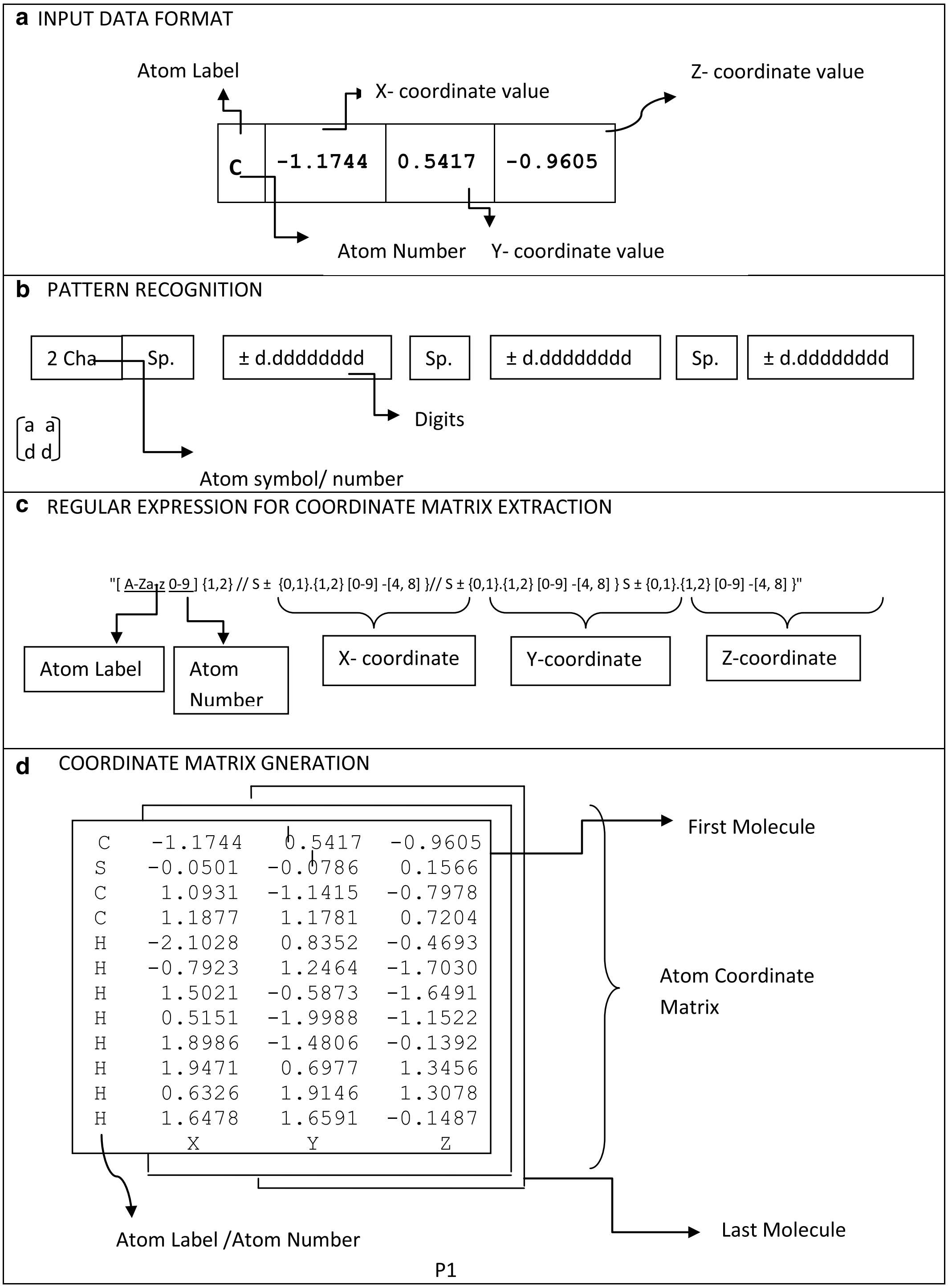
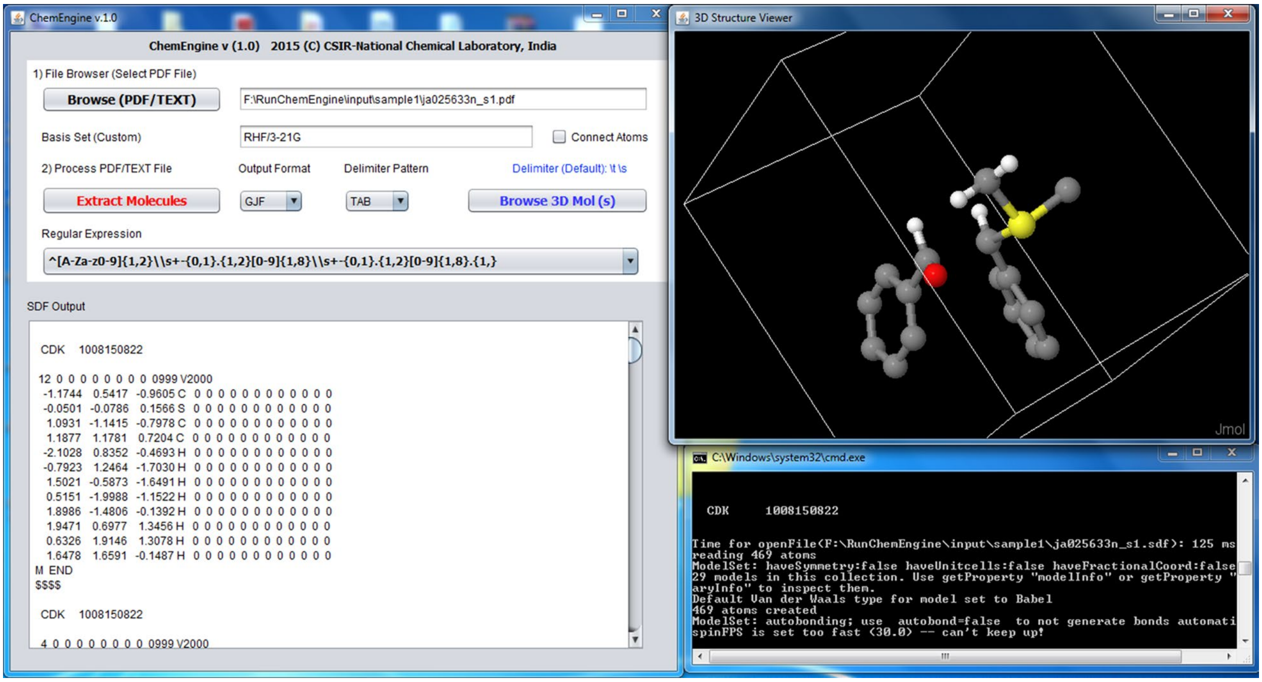

Journal Supporting Material
PDF Parser
Text Data
Pattern Recognition
Extract 3D Coordinates
Build Molecule
Error correction and verify molecules
Bond length, bond angle, dihedral angle
Save Coordinates in .gjf file
G09
Build Bond Matrix
Open babel, CDK descriptors
Single Point Energy
Compute Data
Convert to other standard formats
RHF
Other Basis Sets
Energy Profile
Frequencies
UV-Vis
NMR

Fig. 2 Computational steps workflow for extracting re-computable molecular structures from PDF articles as implemented in ChemEngine



C	-1.1744	0.5417	-0.9605
S	-0.0501	-0.0786	0.1566
C	1.0931	-1.1415	-0.7978
C	1.1877	1.1781	0.7204
H	-2.1028	0.8352	-0.4693
H	-0.7923	1.2464	-1.7030
H	1.5021	-0.5873	-1.6491
H	0.5151	-1.9988	-1.1522
H	1.8986	-1.4806	-0.1392
H	1.9471	0.6977	1.3456
H	0.6326	1.9146	1.3078
H	1.6478	1.6591	-0.1487
	X	Y	Z

Fig. 3 Detailed illustration of key steps in regular expression based pattern recognition for the generation of coordinate matrix data



ChemEngine v (1.0) 2015 (C) CSIR-National Chemical Laboratory, India

1) File Browser (Select PDF File): F:\RunChemEngine\input\sample1\ja025633n_s1.pdf

Basis Set (Custom): RHF/3-21G

□ Connect Atoms

2) Process PDF/TEXT File

Output Format: GJF

Delimiter Pattern: TAB

Delimiter (Default): \t \s

Regular Expression: ^[A-Za-z0-9]{1,2}\\s+-{0,1}.{1,2}[0-9]{1,8}\\s+-{0,1}.{1,2}[0-9]{1,8}.{1,}

SDF Output

```
 CDK    1008150822

12  0  0  0  0  0  0  0  0  0999 V2000
  -1.1744    0.5417   -0.9605 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.0501   -0.0786    0.1566 S   0  0  0  0  0  0  0  0  0  0  0  0
   1.0931   -1.1415   -0.7978 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.1877    1.1781    0.7204 C   0  0  0  0  0  0  0  0  0  0  0  0
  -2.1028    0.8352   -0.4693 H   0  0  0  0  0  0  0  0  0  0  0  0
  -0.7923    1.2464   -1.7030 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.5021   -0.5873   -1.6491 H   0  0  0  0  0  0  0  0  0  0  0  0
   0.5151   -1.9988   -1.1522 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.8986   -1.4806   -0.1392 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.9471    0.6977    1.3456 H   0  0  0  0  0  0  0  0  0  0  0  0
   0.6326    1.9146    1.3078 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.6478    1.6591   -0.1487 H   0  0  0  0  0  0  0  0  0  0  0  0
M  END
$$$$

 CDK    1008150822

 4  0  0  0  0  0  0  0  0  0999 V2000
```

```
  CDK     1008150822

Time for openFile(F:\RunChemEngine\input\sample1\ja025633n_s1.sdf): 125 ms
reading 469 atoms
ModelSet: haveSymmetry:false haveUnitcells:false haveFractionalCoord:false
29 models in this collection. Use getProperty "modelInfo" or getProperty "
aryInfo" to inspect them.
Default Van der Waals type for model set to Babel
469 atoms created
ModelSet: autobonding; use  autobond=false  to not generate bonds automati
spinFPS is set too fast (30.0) -- can't keep up!
```

Fig. 4 ChemEngine GUI for extracting molecular information from the supplementary file selected by the user. The extracted molecules can be dynamically viewed in a 3D viewer

a provision to specify the delimiter such as tab, comma and space if required. The user can select the desired regular expression to extract the coordinate data. Further, in future a customized regular expression can be incorporated into the system based on a particular journal standards of accepting coordinate data in a PDF file. On clicking the *connect atom* button in the browser window, the connection table for a group of coordinates representing a molecule is created and displayed in the output text area. The molecules thus recreated are stored as GJF and SDF format for future computational use and other database oriented inventory applications.

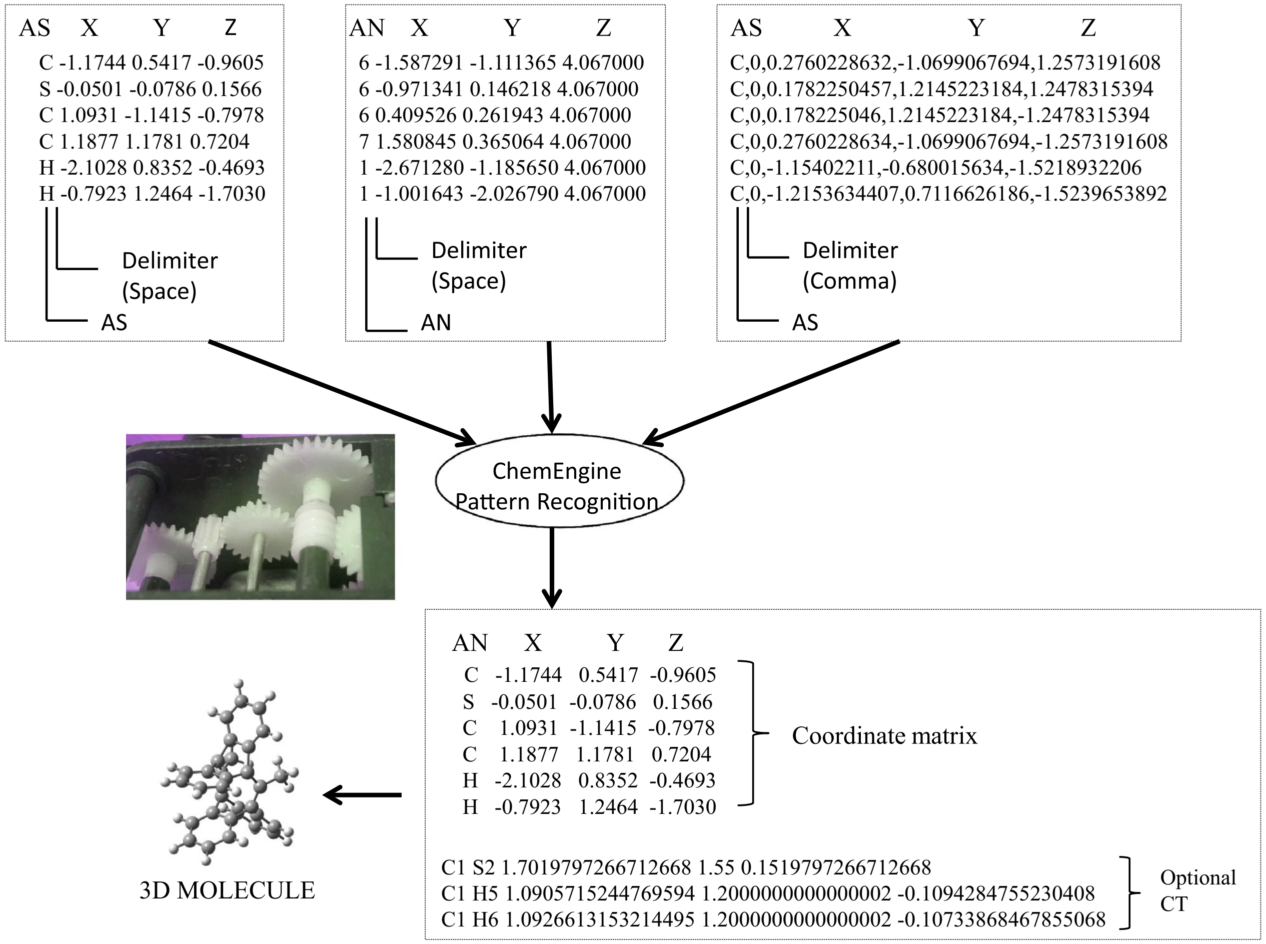
Results and discussion

In general the major problem in processing molecular data stored in PDF files arises due to the non-standard representation of coordinates such as inconsistency in the number of digits appearing after the decimal, interchange of atom type with atomic number in the first column and improper alignment of x, y, z coordinate values. Three case studies have been chosen each dealing with a different representation of coordinate data format in the supplementary information. In the first case, ChemEngine could directly handle the given pdf file and extract the coordinate information. The process in the second case was not straight forward (due to error in PDF to text converter) so the PDF file was first saved as a. txt file externally and then processed to get the desired molecular data. The third case was even more challenging as the molecular coordinates were published in a comma delimiter form (Fig. 5).

Case study 1

The supporting material file was related to reaction modeling research paper describing the mechanistic investigation of epoxide formation from sulfur ylides and aldehydes [32]. The work provided guidelines on stereo-selective synthesis of epoxide ring systems. The computational data included optimized geometries, calculated single point energies, rotational profiles and potential energy surface (PES) generation using standard B3LYP based DFT method. The PDF file was processed to directly extract a.txt file from which patterns were discerned to generate the bond matrix data. For a complete list of coordinate data of molecules generated by ChemEngine please refer to Additional file 1. This file can be considered as a standard template for submitting coordinate data of molecules for fast processing of PDF files in future.

An important constraint for generating ready to compute molecules was the non-availability of bond order information in the published coordinates data.



```
AS   X       Y       Z
  C -1.1744 0.5417 -0.9605
  S -0.0501 -0.0786 0.1566
  C 1.0931 -1.1415 -0.7978
  C 1.1877 1.1781 0.7204
  H -2.1028 0.8352 -0.4693
  H -0.7923 1.2464 -1.7030
```

```
AN   X          Y          Z
 6 -1.587291 -1.111365 4.067000
 6 -0.971341 0.146218 4.067000
 6 0.409526 0.261943 4.067000
 7 1.580845 0.365064 4.067000
 1 -2.671280 -1.185650 4.067000
 1 -1.001643 -2.026790 4.067000
```

```
AS          X                 Y                 Z
C,0,0.2760228632,-1.0699067694,1.2573191608
C,0,0.1782250457,1.2145223184,1.2478315394
C,0,0.178225046,1.2145223184,-1.2478315394
C,0,0.2760228634,-1.0699067694,-1.2573191608
C,0,-1.15402211,-0.680015634,-1.5218932206
C,0,-1.2153634407,0.7116626186,-1.5239653892
```

```
AN    X         Y        Z
 C  -1.1744   0.5417  -0.9605
 S  -0.0501  -0.0786   0.1566
 C   1.0931  -1.1415  -0.7978
 C   1.1877   1.1781   0.7204
 H  -2.1028   0.8352  -0.4693
 H  -0.7923   1.2464  -1.7030
```

```
C1 S2 1.7019797266712668 1.55 0.1519797266712668
C1 H5 1.0905715244769594 1.2000000000000002 -0.1094284755230408
C1 H6 1.0926613153214495 1.2000000000000002 -0.10733868467855068
```

Fig. 5 A schematic diagram highlighting the challenges posed by the diverse coordinate formats present in the supplementary table of journal articles selected for this study. ChemEngine identifies the text patterns and processes this information to yield a common generic format of coordinate matrix. Further the bond matrix algorithm implemented in the program generates a bond matrix for the creation of a connection table to generate the ready to compute 3D molecular structure. *AN* atomic number, *AS* atomic symbol, *CT* connection table

Accordingly functionality has been built in ChemEngine for creating a bond matrix i.e. inter-atomic connectivities of a given cluster of atoms, to facilitate its recognition by the program as a molecule. This enables construction of the connection tables of molecules to assist the direct conversion of a PDF file to SDF on the fly. The method accurately retained the original conformations of all the optimized molecules when the extracted atomic coordinates were supplied back to the original program (Additional file 2).

Understanding the atomic (electronic) movements and distances is of paramount importance in transition state modeling studies of organic reactions. Typically the cut-off distance for the presence of a bond is computed as the sum of the covalent radii of the two atoms, but researchers generally prefer to conduct a computationally less intensive QM calculation and determine based on Wiberg bond order as implemented in the QMDFF code [33]. We took into account the interatomic distances of all the elements in periodic table to annotate the bond order between two atoms. The logic implemented in ChemEngine for creating a bond matrix between two atoms A_1 and A_2 in a molecule is schematically represented in Fig. 6. The cut off distance between two vicinal atoms involved in a covalent bond formation was calculated as the sum of atomic radii + a scaling factor of 0.35 Å, any distance higher than this was considered as a non bonding interaction by the program. Likewise all interatomic distance of other atoms were computed to generate bond matrix of a molecule.

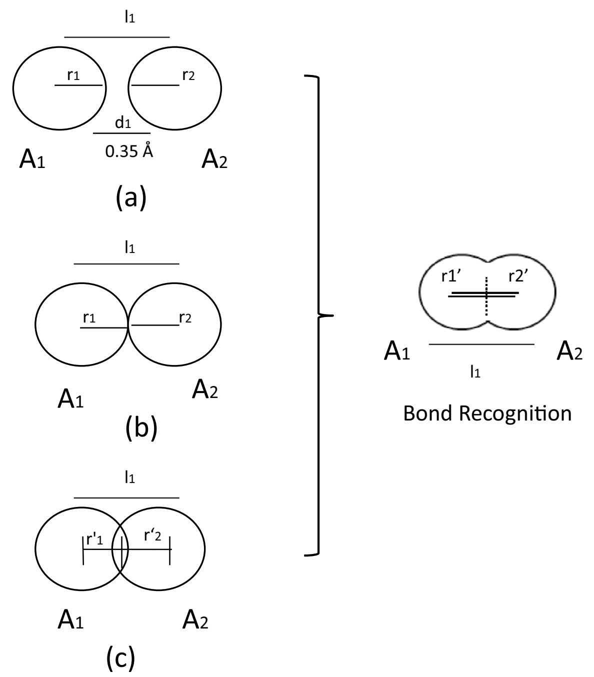


Fig. 6 A schematic depicting the bond recognition logic implemented in ChemEngine v 1.0. **a**–**c** represent three scenarios between two interacting atoms *A1* and *A2*, wherein a bond was considered to be present

To validate our method, the bond matrix for atoms of all the molecules (n = 29) deposited in the supplementary information of the research article was computed and compared with the ones generated by the original software (Gaussian). The values were identical in both the cases. Bond matrix conformation of a representative molecule from this set is shown in Fig. 7 (Bond matrix of few more molecules is shown in Additional file 3). The coordinate data and the computed connectivity information could be used to generate molecules in the SDF and MOL formats.

Case study 2

The work pertains to a well cited paper wherein computational studies were performed on a range of alkenes to gain insights into the mechanistic processes involved in the thiol ene reactions [34] typically classified under click chemistry. In contrast with the previous case study, where the approach was straight forward and an open source pdf reader could be employed to convert pdf to text from the supporting information submitted in a pdf file, in the present case the pdf file was first saved in a plain text format externally and then submitted to ChemEngine for extracting the coordinates. The inadvertent errors in file conversion could be related to compatibility issues associated with various PDF maker programs available on the web.

ChemEngine program could successfully generate the Cartesian coordinates, bond matrix and non molecular data of all the reported molecules (Table 1). Due to the pagination problem in the original PDF document, only few structures partially failed (few atoms carry forward to next molecule) by the program. This pagination issue was later addressed by molecular block identifier—a simple subroutine with the help of which the program could correctly identify molecules reported in a document.

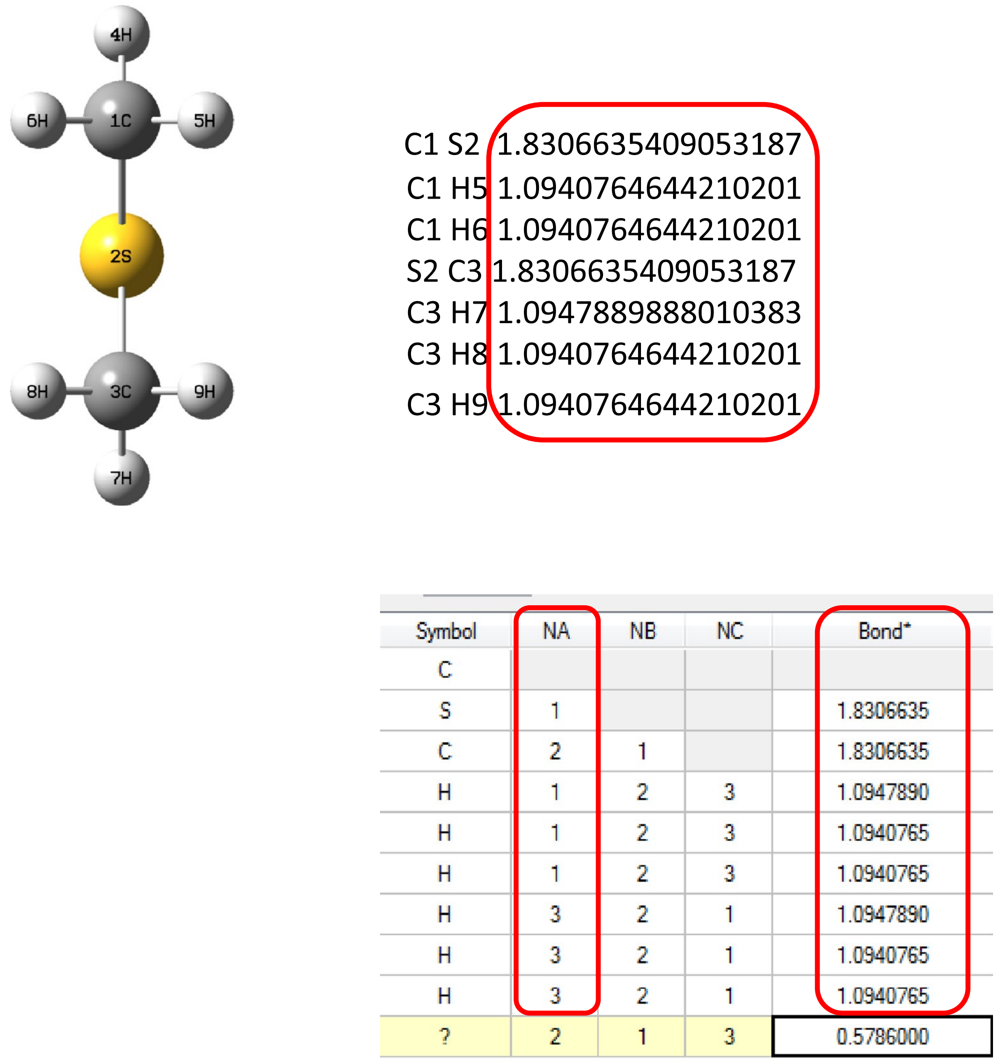
To establish the reusability of the molecular conformations extracted from ChemEngine, we computed their single point energies. RHF (restricted Hartree–Fock) method, a central starting point for multi electron systems was employed for quick computations. The resultant energy values were plotted against the original reported free energies derived from CBS-QB3 (Complete Basis Set), a computationally intensive composite method for yielding very accurate energies. Both the energies were in close agreement (R2 = 0.998) despite the choice of different methods (Fig. 8).

Case study 3

The computational work reported involved Cope rearrangement transition states using the DFT method to compute electronic energies for various substituted allyl derivatives [35]. This example consisted of a PDF file wherein the coordinates data was submitted in a comma separated format in the supplementary file. The code implemented in ChemEngine was modified to parse any coordinates data interspersed with delimiters such as comma, tab or space in a PDF file. The results of all three case studies are summarized in Table 2 which prove the robustness and efficiency of the ChemEngine program in recognizing patterns and developing regular expression for the typical cases dealt (Additional files 4, 5).

Case study 4

In order to increase the scope of this work to handle several hundred PDF files to harvest truly computable molecular data, that are buried in PDF files we have implemented a default option in ChemEngine to harvest atomic co-ordinate data mixed with images (spectral data, barcode images, experimental data, molecular description and other computed data) and successfully tested with several PDF files to regenerate molecular files without any errors [36].



C1	S2	1.8306635409053187
C1	H5	1.0940764644210201
C1	H6	1.0940764644210201
S2	C3	1.8306635409053187
C3	H7	1.0947889888010383
C3	H8	1.0940764644210201
C3	H9	1.0940764644210201

Symbol	NA	NB	NC	Bond*
C				
S	1			1.8306635
C	2	1		1.8306635
H	1	2	3	1.0947890
H	1	2	3	1.0940765
H	1	2	3	1.0940765
H	3	2	1	1.0947890
H	3	2	1	1.0940765
H	3	2	1	1.0940765
?	2	1	3	0.5786000

Fig. 7 Interatomic bond distances reproduced using ChemEngine by harvesting the 3D coordinate structural data from the pdf file (Mol ID 29, Dimethyl sulfide). The bond characteristics were identical with those generated by the original program (GaussView)

In the present work we process the molecules and transform them into SDF format that is mostly compatible with commercial packages thus saving time and computational effort. The compute once and use many times approach will help the readers to access the original input files even after passage of time. It is pertinent to mention here that the biological sciences and bioinformatics community follow a standard representation of molecular coordinates in the PDB file format which is a database compliant format instead of a PDF format thus securing an easy access and exchange of information. Extracting coordinates of protein molecule from a PDF file, assuming an average protein size of over 2,00,000 atoms would have been indeed a truly challenging task. However with the aid of ChemEngine customized with additional atomic co-ordinate pattern recognition modules, now it is possible to harvest any molecular data from PDF format. With the advent of 3D structure repositories and several free academic sites, data storage is no longer a major issue, the ready to compute molecules can be deposited and maintained to avoid duplication of computational efforts. Till such a global archival norm is achieved, it is suggested that the chemical community should maintain a standard and consistent representation

Table 1 Sample data from three output files viz. bond matrix, coordinate data and non-molecular text generated by ChemEngine in the second case study related to thiol ene click chemistry. Data displayed here has been truncated for brevity (n = 115)

Bond matrix data	Coordinates Data	Non- Molecular Data
0:---- Mol_0 1 C1 H2 1.0901109643921576 1.2000000000000002 -0.10988903560784258 Mol_0 2 C1 H3 1.0892309436382168 1.2000000000000002 -0.11076905636178336 Mol_0 3 C1 H4 1.0892309436382168 1.2000000000000002 -0.11076905636178336 Mol_0 4 C1 S5 1.831868 1.55 0.281868 Mol_0 5 S5 H6 1.3438373386872386 1.4100000000000001 -0.0661626613127615 1:---- Mol_0 1 C1 H2 1.0967060354908238 1.2000000000000002 -0.10329396450917638 Mol_0 2 C1 H3 1.0907281685566759 1.2000000000000002 -0.10927183144332431 Mol_0 3 C1 H4 1.0907540968948959 1.2000000000000002 -0.10924590310510429 Mol_0 4 C1 S5 1.8058019435159547 1.55 0.25580194351595464 2:---- Mol_0 1 C1 H2 1.0862053681380885 1.2000000000000002 -0.11379463186191163 Mol_0 2 C1 H3 1.0841699092264088 1.2000000000000002 -0.11583009077359141 Mol_0 3 C1 C4 1.3293846809276089 1.34 - 0.010615319072391216 Mol_0 4 C4 H5 1.0888505631306806 1.2000000000000002 -0.11114943686931955 Mol_0 5 C4 C6 1.5004337593112866 1.34 0.16043375931128656 Mol_0 6 C6 H7 1.095743612037506 1.2000000000000002 -0.10425638796249426	C -0.04781100 1.16216400 0.00000000 H -1.09556300 1.46309200 0.00000000 H 0.43082600 1.55738100 0.89506100 H 0.43082600 1.55738100 - 0.89506100 S -0.04781100 -0.66970400 0.00000000 H 1.28575000 -0.83557700 0.00000000 C -1.11122700 0.00005600 - 0.00880200 H -1.42403800 -0.00270000 1.04234300 H -1.51094200 0.90050300 - 0.47689500 H -1.51064400 -0.89830400 - 0.48120000 S 0.69456200 0.00001000 - 0.00196500 C -1.28038600 0.22044600 - 0.00000100 H -1.30140400 1.30644800 - 0.00003900 H -2.23896200 -0.28606800 0.00010900 C -0.13464400 -0.45374900 - 0.00003700 H -0.16675400 -1.54212600 0.00001400	S1 SUPPORTING INFORMATION Thiol-Ene Click Chemistry: Computational and Kinetic Analysis of the Influence of Alkene Functionality. Brian H. Northrop* and Roderick N. Coffey Department of Chemistry Wesleyan University, Middletown, Connecticut 06459. TABLE OF CONTENTS I. Electron density maps of alkenes 2-13 S2 II. Relative energetics of exo and endo propagation and S3 chain-transfer steps for norbornene III. Computed and experimental ionization potentials and electron affinities S4 IV. ?HCTï¿½ versus Cï¿½H and Sï¿½H transition state bond distances S4 V. Carbon-centered radical intermediate radical stabilization energies S5 VI. Calculated transition state entropies, pre-exponential factors, S5 and activation energies VII. Calculated versus experimental rates of propagation and chain-transfer

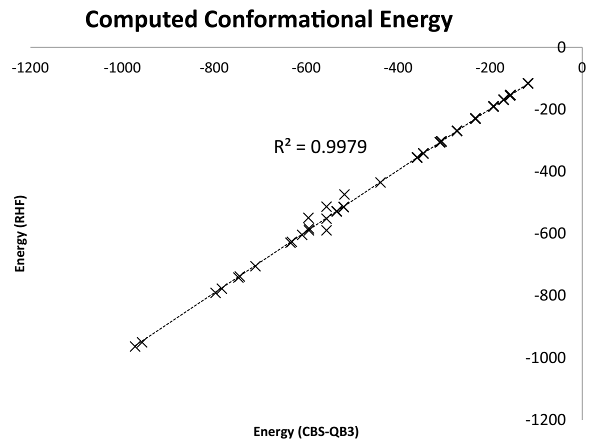


Fig. 8 A comparative plot of single point energies of molecules extracted from coordinate data related to case study 2. The values are in agreement with the original computed data. *CBS* Complete Basis Set, *RHF* restricted Hartree–Fock

of chemical structure data in the electronic supplementary files in native format or standard data format to facilitate the re-usability among the scientific community.

Conclusion

Supplementary information of primary literature deposited with journals is a rich reservoir of peer reviewed molecular data which will be more valuable if available for further reuse. An application ChemEngine presented here selectively extracts the 3D structure from coordinate information present along with inadvertently introduced noisy data present in PDF files. This approach can obviate to some extent the loss of chemical data while at the same time conserve the memory and storage space required at the journal site. The methodology exemplified here will enable molecule mining in semantic context and ensure maximum reuse of the valuable data by

Table 2 Details of the three case studies representing the diversity of coordinate molecular data in supplementary material handled by ChemEngine

Entry	Case study	N = molecules	Regular expression pattern	Format and delimiter
1	Epoxide formation from sulfur ylides and aldehydes	29	^[A-Za-z0-9]{1,2}\\s+-{0,1}.{1,2}[0-9]{1,8}\\s+-{0,1}.{1,2}[0-9]{1,8}.{1,}	PDF Space
2	Thiol ene click chemistry	115	^[A-Za-z0-9]{1,2}\\s+-{0,1}.{1,2}[0-9]{1,8}\\s+-{0,1}.{1,2}[0-9]{1,8}.{1,}	Text Space
3	Design of tetra(arenediyl)bis(allyl) derivatives for cope rearrangement transition states	55	^[A-Za-z0-9]{1,2}\\,[0]{0,1}[\\,]{0,1}-{0,1}.{1,2}[0-9]{1,10}\\,-{0,1}.{1,2}[0-9]{1,10}.{1,}	PDF Comma

interested readers thereby enhancing the citations of the authors. Further the application can be seamlessly integrated to enable a high throughput molecular computing automated workflow.

Additional files

Additional file 1. Coordinate data of 29 molecules obtained as output from ChemEngine in the first case study pertaining to formation of epoxides from sulfur ylides and aldehydes.

Additional file 2. Recreated 3D geometry optimized structures of 29 molecules as visualized in the original program (Gauss View).

Additional file 3. A comparative table consisting of interatomic bond distances computed via ChemEngine and the commercial software package. This material is available free of charge via the Internet at http://pubs.acs.org.

Additional file 4. Sample input file containing co-ordinates of molecules separated by comma delimiter.

Additional file 5. Instruction for compilation of chemengine source code available online and operation manual.

Abbreviations

PDF: printable document format; ORTEP: oak ridge thermal ellipsoid plot; SDF: structure data format; QMDFF: quantum mechanically derived force field.

Authors' contributions

MK conceived the idea and developed the software, RV validated the methodology, tested the application and prepared the manuscript. Both authors read and approved the final manuscript.

Author details

[1] Present Address: Chemical Engineering and Process Development (CEPD), CSIR-National Chemical Laboratory, Pashan Road, Pune, Maharastra 411008, India. [2] MIT School of Bioengineering Sciences and Research, ADT (Art, Design and Technology) University, Loni Kalbhor, Pune, Maharashtra 412201, India.

Acknowledgements

MK thanks director CSIR NCL for providing infrastructure and support. The financial support received from GENESIS (BSC0121) and INSPIRE (CSC0107) under 12FYP projects is duly acknowledged. RV thanks DST, New Delhi for award of a fellowship. We also thank reviewers for valuable comments to improvise the tool with default options that increased the scope of ChemEngine to handle several hundred PDF files to generate molecular structures with high accuracy.

Competing interests

The authors declare that they have no competing interests.

Received: 2 March 2016 Accepted: 18 October 2016
Published online: 29 December 2016

References

1. Karthikeyan M, Vyas R (2015) Role of open source tools and resources in virtual screening for drug discovery. Comb Chem High Throughput Screen 18(6):528–543
2. Blurock E (1995) Reaction: system for modeling chemical reactions. J Chem Inf Model 35(3):607–616
3. Dolata D, Spina D, Stahl M (1996) Conformational searching and modeling of transition states. J Chem Inf Model 36(2):228–230
4. Aziz H, Gao J, Maropoulos P, Cheung W (2005) Open standard, open source and peer-to-peer tools and methods for collaborative product development. Comput Ind 56(3):260–271
5. Karthikeyan M, Vyas R (2015) Role of open source tools and resources in virtual screening for drug discovery. Comb Chem High Throughput Screen 18(6):528–543
6. Gilbert I (2013) Drug discovery for neglected diseases: molecular target-based and phenotypic approaches. J Med Chem 56(20):7719–7726
7. Ryall K, Tan A (2015) Systems biology approaches for advancing the discovery of effective drug combinations. J Cheminform 7(1):7
8. Postma G, van Bakel B, Kateman G (1996) Automatic extraction of analytical chemical information. System description, inventory of tasks and problems, and preliminary results. J Chem Inf Model 36(4):770–785
9. Karthikeyan M, Bender A (2005) Encoding and decoding graphical chemical structures as two-dimensional (PDF417) barcodes. J Chem Inf Model 45(3):572–580
10. Murray-Rust P, Mitchell J, Rzepa H (2005) Chemistry in bioinformatics. BMC Bioinform 6(1):141–144
11. Murray-Rust P, Mitchell J, Rzepa H (2005) Communication and re-use of chemical information in bioscience. BMC Bioinform 6(1):180–195
12. Guha R, Howard M, Hutchison G, Murray-Rust P, Rzepa H, Steinbeck C et al (2006) The blue obelisk interoperability in chemical informatics. J Chem Inf Model 46(3):991–998
13. https://jchempaint.github.io/. Accessed 27 Sept 2016
14. http://sourceforge.net/projects/cdk/. Accessed 27 Sept 2016
15. Steinbeck C, Krause S, Kuhn S (2003) NMRShiftDB—constructing a free chemical information system with open-source components. J Chem Inf Model 43(6):1733–1739
16. http://www.ccdc.cam.ac.uk/. Accessed 27 Sept 2016
17. Fourches D, Muratov E, Tropsha A (2015) Curation of chemogenomics data. Nat Chem Biol 11(8):535
18. Harvey MJ, Mason NJ, McLean A, Murray-Rust P, Rzepa HS, Stewart JJP (2015) Standards-based curation of a decade-old digital repository dataset of molecular information. J Cheminform 7:43
19. http://opsin.ch.cam.ac.uk. Accessed 27 Sept 2016
20. O'Donnell T (2009) Design and use of relational databases in chemistry. CRC Press, Boca Raton
21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3945766/pdf/f1000research-3-3897.pdf
22. Richard A, Williams C (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. Mutat Res 499(1):27–52
23. Karthikeyan M, Krishnan S, Pandey A, Bender A (2006) Harvesting chemical information from the internet using a distributed approach: ChemXtreme. J Chem Inf Model 46(2):452–461
24. http://www.chemaxon.com/. Accessed 27 Sept 2016
25. Karthikeyan M (2011) Automatic harvesting of molecular information raster graphics. US Patent Appl 14/241285
26. http://cactus.nci.nih.gov/osra/. Accessed 27th Sept 2016

27. Gkoutos G, Rzepa H, Clark R, Adjei O, Johal H (2003) Chemical machine vision: automated extraction of chemical metadata from raster images. J Chem Inf Model 43(5):1342–1355
28. Ibison P, Jacquot M, Kam F, Neville A, Simpson R, Tonnelier C et al (1993) Chemical literature data extraction: the CLiDE Project. J Chem Inf Model 33(3):338–344
29. Feldman R, Sanger J (2007) The text mining handbook. Cambridge University Press, Cambridge
30. Karthikeyan M, Pandit Y, Pandit D, Vyas R (2015) MegaMiner: a tool for lead identification through text mining using chemoinformatics tools and cloud computing environment. Comb Chem High Throughput Screen 18(6):591–603
31. http://www.gaussian.com/. Accessed 27 Sept 2016
32. Aggarwal V, Harvey J, Richardson J (2002) Unraveling the mechanism of epoxide formation from sulfur ylides and aldehydes. J Am Chem Soc 124(20):5747–5756
33. Grimme S (2014) A general quantum mechanically derived force field (QMDFF) for molecules and condensed phase simulations. J Chem Theory Comput 10(10):4497–4514
34. Northrop B, Coffey R (2012) Thiol ene click chemistry: computational and kinetic analysis of the influence of alkene functionality. J Am Chem Soc 134(33):13804–13817
35. Salvatella L (2015) Theoretical design of tetra(arenediyl)bis(allyl) derivatives as model compounds for Cope rearrangement transition states. RSC Adv 5(15):11494–11497
36. https://sourceforge.net/projects/chemengine/files/?source=navbar. Accessed 27 Sept 2016