Catalysis Today xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Catalysis Today

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

Geographical information system software as in-house chemical indexing database for catalyst screening of alkene metathesis catalysts

Jean I. du Toit[a,*], Marié J. du Toit[b], Cornelia G.C.E. van Sittert[a], Hermanus C.M. Vosloo[a]

[a] Research Focus Area for Chemical Resource Beneficiation, Catalysis and Synthesis Research Group, North-West University, Potchefstroom, 2520, South Africa
[b] Unit for Environmental Sciences and Management, North-West University, Potchefstroom, 2520, South Africa

ARTICLE INFO

Keywords:
Geographical information system
Chemical indexing database
Catalyst screening
Attribute tables
Visualization

ABSTRACT

Large amounts of chemical information can be indexed on various databases. If these databases store information in an appropriate format for specific search queries, we can perform a catalyst screening from intelligent searches. However, little to no database applications are flexible enough to allow for simultaneous screening and comparable visualizations. Therefore, for setting up a flexible in-house database, we propose the use of geographical information system (GIS) software. GIS includes an attribute table that can be populated with text and numerical information. From this table, we could selectively screen our data as well as visualize the different values on an intelligently created grid pattern. Consequently, 11 types of alkene metathesis catalysts together with 9 Grubbs-type catalysts were investigated using GIS techniques. GIS proved to be an efficient and versatile tool whereby quick and successful catalyst screening could be done.

1. Introduction

Through experimental and computational work in chemistry vast amounts of data are created daily. In Fig. 1 we see a metaphorical illustration of these large datasets as an “iceberg” of data. These data icebergs typically include calculated molecular properties (e.g. bond lengths, energies, bond order, etc.) and experimental data (e.g. GC results, IR results, etc.). Visually, we can only handle the tip of the iceberg. The rest of the iceberg or data forms an imperceptible mountain below the surface. Therefore, it is difficult for the researcher to efficiently digest this data due to the sheer amount that needs to be processed.

A possible solution to process the data is by employing a data-orientated research approach called informatics; which in the field of chemistry is called chemoinformatics [1]. Chemoinformatics “is the mixing of information to transform data into information and information into knowledge for the sole purpose to make better decisions faster in the identification and development of new chemicals for specific purposes” [2]. Chemoinformatics, therefore, enables various techniques with which, for example, catalyst screening can be done. One of these techniques is by using chemical databases [3]. These databases (i.e. PubChem [4]) are a way to index chemical information in text and numerical formats [3]. Those formats then allow the researcher to perform various searches from which further investigations can be done. The search method then depends on the type of query you have. For instance, if the query is based on the structure-activity relationships between molecules, a quantitative structure-activity relationship (QSAR) analysis can be done to extract the information [3]. Therefore, “these common tools enable exponential knowledge discovery within exponentially growing data collections” [5].

These tools, however, do not allow for the flexibility of a one-stop solution for extraction and comparative visualization. Various techniques and software applications need to be used in combination. Consequently, if you can set up an in-house database of all experimental and computational data with suitable software, the screening can be done together with comparative visualizations. Therefore, for setting up a flexible in-house database, we propose the use of geographical information system (GIS) [6] software [7]. According to the Environmental Systems Research Institute (ESRI) [7]:

> A geographic information system (GIS) is a framework for gathering, managing, and analyzing data. Rooted in the science of geography, GIS integrates many types of data. It analyzes spatial location and organizes layers of information into visualizations using maps and 3D scenes. With this unique capability, GIS reveals deeper insights into data, such as patterns, relationships, and situations—helping users make smarter decisions.

The key to managing GIS information is an attribute table [7] that

* Corresponding author.
E-mail address: 12317624@nwu.ac.za (J.I. du Toit).

https://doi.org/10.1016/j.cattod.2019.01.032
Received 31 August 2018; Received in revised form 7 December 2018; Accepted 12 January 2019

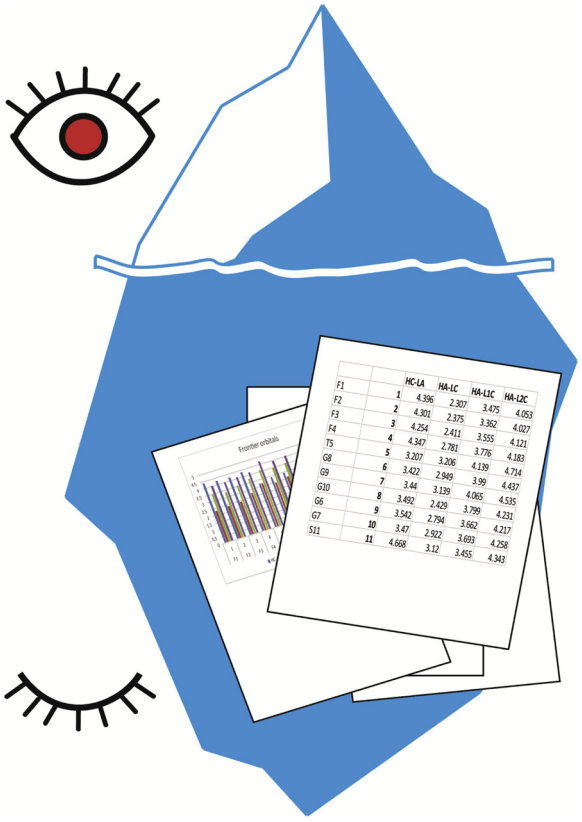

Fig. 1. Iceberg of data.

can be populated with text and numerical information. An attribute table is similar in appearance to any typical spreadsheet with columns and rows, but each row is specifically linked to a spatial object on a map such as a line representing a road. From an attribute table, linked to for example an user created spatial grid pattern, it would be possible to selectively screen data populated in the attribute table as well as visualize the different values (linked to individual grid blocks representing e.g. catalysts) through the graphic user interface of the GIS software. By using one programme that can handle both advanced search queries and multiple visualization outputs, a researcher can not only greatly reduce the time in doing chemical queries but also have an excellent way of generating graphic output that both informs and satisfies.

The purpose of this paper is two-fold: firstly, to demonstrate the capability of GIS as an in-house chemical database, and secondly, as a case-study, to screen metal complexes as catalysts for alkene metathesis. Therefore, the main aim was to illustrate the advantage of using GIS as a screening technique by comparing the statistically generated data for 11 metal carbene complexes from our previous study [8] with the GIS screening techniques. The secondary aim was then to include a short application to show the implementation of the GIS screening. This application was done by expanding the selection to include more Grubbs-type metal carbenes.

2. Materials and method

2.1. GIS software

Several GIS software packages exist with which the data analyses and visualization described below can be done. Popular software packages include the commercially available ArcGIS from ESRI (https://www.esri.com/en-us/home) and the open source software QGIS (https://www.qgis.org/en/site/). The methodology described in this paper explain the steps using ArcGIS software specifically.

The main data input format of GIS is spatial data. Spatial data can be defined as “information about the locations and shapes of geographic features and the relationships between them, usually stored as coordinates and topology” [7]. However, by seeing chemical information as “spatial data” linked to specific items, i.e. molecules, we can implement GIS in chemistry applications. Real world objects created in GIS are called features [7], a set of features is called a layer, and each layer has an automatically generated attribute table associated with it. An attribute table is defined as “a database or tabular file containing information about a set of geographic features, usually arranged so that each row represents a feature and each column represents one feature attribute. In GIS, attribute tables are often joined or related to spatial data layers, and the attribute values they contain can be used to find, query, and symbolize features or raster cells” [7]. The automatically generated attribute tables are editable allowing the addition of columns to the table.

Tables can also be imported into GIS as standalone objects and used for analyses, however, to use the visualization capabilities of GIS, a feature layer must be created with its associated table. We created spatial features to represent chemical objects, e.g. elements or catalysts, in order to utilize the visualization potential of the mapping software. The chemical 'features' can be visualized in any shape or form; logical choices include grids arranged according to some underlying pattern or order, e.g. for elements their position in the periodic table or in the case of catalysts one of their characteristics such as ligand substructures. Each grid block would then represent a specific chemical object. Illustrations of visualization options are displayed in Fig. 2. for example, illustration **a** is a typical grid where each cell depicts a different item or molecule. Secondly, in **b** and **c**, disconnected or connected cells are drawn according to any predetermined order. Finally, in **d** and **e**, any shape can be created all linked to a row in the associated feature attribute table. Therefore, in the illustration, the green, orange and red areas can all represent the same chemical objects; they are just created in different visual forms. For the case study in this paper, the attribute table rows represent different metal carbene catalysts. These catalysts are organised and visualized according to the similarity or difference in their substructures as determined by their ligands.

All screening and visualizations were done with the associated GIS software applications from the ESRI suite [7]. For this paper, ArcGIS 10.2 for Desktop [9] was utilized to do the analyses. ArcGIS Desktop consists of several installed applications of which ArcMap is the main package with which map-based analyses and editing is done. Furthermore, another application called ArcScene, specializing in 3D visualization and analysis were also used in this paper. In ArcMap, we created a new feature using the Create Fishnet Tool in the data management toolset. This tool creates a user-customized grid with any number of

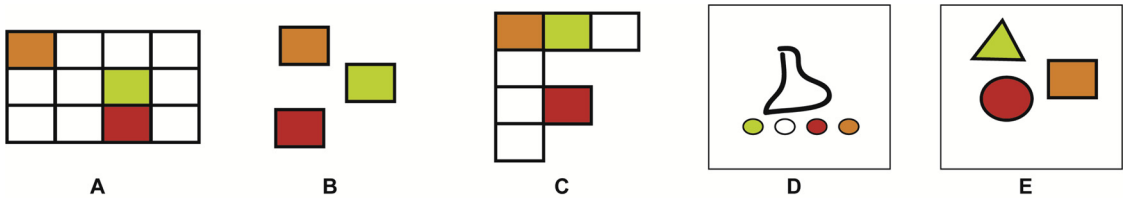

Fig. 2. Examples of GIS visualization forms.