



ugr

Universidad
de Granada

TRABAJO FIN DE GRADO
INGENIERÍA INFORMÁTICA

Detección de errores en bases de datos químicas

Autor

Jesús Navarro Merino

Directora

Rocío Celeste Romero Zaliz



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN

Granada, julio de 2023

Detección de errores en bases de datos químicas

Jesús Navarro Merino

Palabras clave: palabra_clave1, palabra_clave2, palabra_clave3,

Resumen

Poner aquí el resumen.

Project Title: Project Subtitle

First name, Family name (student)

Keywords: Keyword1, Keyword2, Keyword3,

Abstract

Write here the abstract in English.

Yo, **Jesús Navarro Merino**, alumno de la titulación Grado en Ingeniería Informática de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 15429457E, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Jesús Navarro Merino

Granada a X de MES de 201 .

Dña. **Rocío Celeste Romero Zaliz**, Profesora del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado *Detección de errores en bases de datos químicas*, ha sido realizado bajo su supervisión por **Jesús Navarro Merino**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a X de mes de 201 .

La directora:

Rocío Celeste Romero Zaliz

Agradecimientos

Poner aquí agradecimientos...

Índice general

1. Introducción	1
1.1. Motivación y objetivos	2
1.2. Objetivos	7
1.3. Estructura de la memoria	7
2. Estado del arte y fundamentos teóricos	9
3. Gestión y Planificación del proyecto	11
3.1. Metodología	11
3.2. Gestión de recursos	11
3.2.1. Recursos humanos	11
3.2.2. Recursos materiales	11
3.2.3. Recursos software	11
3.3. Gestión de costes	11
3.3.1. Coste de recursos humanos	11
3.3.2. Coste hardware	11
3.3.3. Otros costes	11
3.3.4. Presupuesto final	11
3.4. Análisis de riesgos	11
4. Experimentación	13
A. Tabla comparativa del set completo de moléculas	17

Índice de figuras

1.1. Distintas cadenas SMILES válidas para el 1-methyl-3-bromo- ciclohexeno. (a) Considera el ciclo como la rama principal y el bromo como ramificación. (b) Hace el recorrido que marca la flecha, dejando parte del ciclo como una ramificación. Imagen extraída de [3]	3
---	---

Índice de tablas

1.1. Códigos SMILES y sus representaciones visuales según Sigma-Aldrich	5
1.2. Códigos SMILES y sus representaciones visuales según SciFinder	6
A.1. Tabla de índices con los nombres de las moléculas de la Tabla A.2	17
A.2. Tabla extendida para el set de datos de 30 moléculas. Contiene la cadena SMILES extraída de Sigma-Aldrich (SA), la cadena SMILES extraída de SciFinder (SF), y las imágenes de las respectivas bases de datos (SA y SF)	19

Capítulo 1

Introducción

La Química estudia la composición y estructura de la materia, sus propiedades y transformaciones. Estudia las sustancias, la energía y sus cambios durante las reacciones. Desde que se tienen registros, la química ha sido fundamental para el desarrollo de la humanidad, ya que ha permitido la producción de materiales, alimentos, medicamentos y energía, entre otros. Esto ha sido un proceso lento y exhaustivo a través de la experimentación. Por ejemplo, en 1881, Beilstein publica su Manual de Química Orgánica, que recogía 15000 compuestos orgánicos con sus propiedades [2]. Conforme la química se iba expandiendo, también lo hacía el volumen de datos que se generaban, siendo cada vez mas frecuentes preguntas como "¿alguien habrá sintetizado ya este compuesto?" [9]

Eventualmente, hace unas cuantas décadas, se pensó que la cantidad de información que cada químico por su cuenta había acumulado, se podía compartir y hacer accesible a la comunidad científica a través de su almacenamiento en bases de datos [4]. Con el desarrollo de técnicas de manipulación y tratamiento de esos datos surgió el término *chemoinformatics*.

Las *chemoinformatics* han cobrado gran importancia en los últimos años debido al aumento exponencial de datos experimentales generados en la investigación biomédica y química, y a la necesidad de manejar y analizar esta información de manera eficiente. Esta disciplina ha sido influenciada por diversas áreas, como la química, matemáticas, estadística, biología y ciencias de la computación entre otras. Al parecer, su origen se remonta a la década de 1940, habiendo ya algunas investigaciones en el área, pero el término 'chemoinformatics' como tal se lleva utilizado más bien poco (1998) [5]. Como tal, aun no hay un acuerdo en cuanto a su definición, seguramente por su carácter interdisciplinar, ni si quiera en cómo deletrearlo, pudiendo aparecer también como *cheminformatics*, *chemical informatics*, *chemi-informatics*, y *molecular informatics* entre otras [5, 7]. En la literatura se discuten varias interpretaciones sobre su definición, unas más precisas y

otras más generales: [5, 6, 7, 4]

La mezcla de recursos de información para transformar datos en información, y la información en conocimiento, con el fin de tomar decisiones más rápidas y efectivas en la identificación y optimización de fármacos [Brown 1998]

Chem(o)informatics es un término genérico que abarca el diseño, la creación, la organización, la gestión, la recuperación, el análisis, la difusión, la visualización y el uso de la información química. [G. Paris 1999]

La aplicación de métodos informáticos para resolver problemas de química [J. Gasteiger and T. Engel 2006]

A pesar de ello, son a día de hoy un componente esencial en el descubrimiento de sustancias químicas; sin duda es un área en constante evolución y su importancia solo aumentará en los próximos años, tanto en el descubrimiento de fármacos —que es como originariamente surgió y donde más impacto tiene en la sociedad— como en otros campos de la química.

Una herramienta también de vital importancia en este ámbito son los sistemas de representación lineal. Surgieron a medida que la química y la tecnología computacional avanzaban, y nos permiten codificar moléculas para su análisis y almacenamiento en bases de datos. En el siglo XIX, se desarrollaron varias formas de representación visual de moléculas, como las fórmulas estructurales que permitieron a los químicos dibujar y visualizar moléculas de manera más efectiva. Sin embargo, estas formas de representación no son adecuadas para su uso en la computación, ya que no son fácilmente legibles para los programas informáticos. Nosotros, los humanos, cuando vemos una estructura molecular dibujada la entendemos directamente, obtenemos una visión global de los símbolos que representan los enlaces y la distribución espacial de los átomos que la componen, pero los computadores no tienen esa facilidad. Por ello, se desarrollaron sistemas de notación lineal que permitían describir de manera más precisa y eficiente la estructura molecular, trabajando con tipos de datos sencillos, cadenas de caracteres.

1.1. Motivación y objetivos

Los formatos de notación lineal llevan siendo un tema de interés e investigación para los científicos desde mediados del siglo 19, evolucionando poco a poco y desarrollándose nuevas notaciones en función de las necesidades —principalmente computacionales— del momento y las limitaciones que se iban descubriendo [1]. En la actualidad, existen varias representaciones lineales, siendo las más usadas SMILES, InChI, y SELFIES [11]. Como

comenté antes, una forma muy potente de representar moléculas y compuestos químicos es mediante cadenas strings, y de esto justamente se encargan las representaciones lineales: traducir una molécula, con sus átomos, enlaces entre ellos, ciclos y otras propiedades características, en una cadena string que la represente, y que la máquina y los propios químicos puedan entender. Sin embargo, hay diferencias notables entre las representaciones, tanto en la sintaxis de las cadenas que se generan como en las aplicaciones que se le puede dar a cada una de ellas.

SMILES, ideada por David Weininger, sale a la luz en 1988 satisfaciendo con creces las necesidades de procesamiento de información química que había, desbancando a la representación estandarizada del momento, Wiswesser Line Notation (WLN). Desde ese entonces SMILES se convirtió —y sigue siendo a día de hoy— en el estándar de representación lineal, ya que permite describir estructuras moleculares de una forma sencilla en un formato fácil de leer, lo que ha hecho que sea una herramienta popular en la química computacional, siendo la más usada entre investigadores y químicos. Pese a esto, SMILES tiene dos grandes inconvenientes: una misma molécula puede escribirse con varias cadenas SMILES distintas válidas, es decir, tiene sinónimos (Figura 1.1); y no es robusto ni sintáctica ni semánticamente. En este sentido se podría generar un string que no represente una molécula válida, como lo es CC(CCCC, el cual tiene un paréntesis sin cerrar (lo que implica que no se delimita cuándo acaba la rama). O generar una molécula que no sea químicamente viable como CO=CC, que muestra un átomo de oxígeno neutro formando tres enlaces (superando el límite de enlaces covalentes que un oxígeno neutro puede tener) [11].

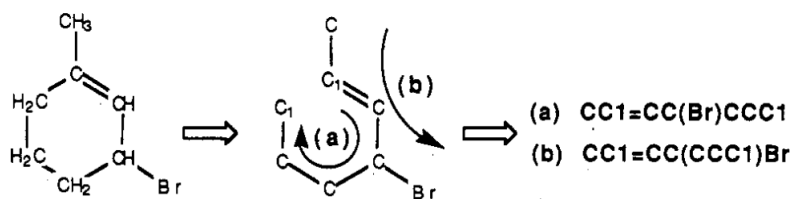


Figura 1.1: Distintas cadenas SMILES válidas para el 1-methyl-3-bromociclohexeno. (a) Considera el ciclo como la rama principal y el bromo como ramificación. (b) Hace el recorrido que marca la flecha, dejando parte del ciclo como una ramificación. Imagen extraída de [3]

Esto tiene especial relevancia en el ámbito del Machine Learning (ML). Aunque se sale del alcance de este trabajo, uno de los grandes objetivos de la química computacional es la creación o diseño de nuevas moléculas. Se podrían crear modelos de ML o redes neuronales capaces de generar moléculas ficticias válidas, para posteriormente ver sus propiedades, valorarlas energéticamente para ver cuán estables son, y estudiar su viabilidad en

distintas aplicaciones, entre otras cosas. SMILES dificulta esta tarea, y por ello, aparece en 2020, SELFIES (SELF-referencIng Embedded Strings), una nueva representación lineal 100 % robusta, muy usada actualmente para modelos generativos. Ver [11, 10] para más detalles de cómo soluciona los problemas de robustez y otras características de la representación. SELFIES es relativamente reciente y continuamente está ampliando sus funcionalidades, mejorando su simplicidad y facilidad de uso para el usuario [12]. Aun así, no se termina de instaurar entre la comunidad investigadora. Por último, InChI es creado en 2013 por la IUPAC (International Union of Pure and Applied Chemistry) como un proyecto para estandarizar el proceso de búsqueda de estructuras moleculares entre distintas bases de datos. Esto es porque InChI (International Chemical Identifier) genera una cadena canónica única para cada molécula, de manera que cada molécula tiene una sola representación, y dicha representación solamente hace referencia a esa molécula. La principal desventaja radica en su sintaxis y su estructura jerárquica, haciéndola complicada de leer y utilizar por los humanos. Por esto mismo también, no es la mejor opción para usar en modelos generativos, pues tiene una serie de reglas y normas gramaticales y aritméticas que son complejas de aplicar al generar moléculas a través de modelos de ML.[8]

Por todo lo anterior, me centraré en la notación SMILES durante el desarrollo de este trabajo. Dicho esto, existen diversas bases de datos en química donde se recoge gran cantidad de información acerca de los compuestos. Entiéndase esto como una colección estructurada y organizada que contiene datos sobre compuestos químicos, sus propiedades y relaciones con otros compuestos. Se utilizan para almacenar y recuperar información sobre moléculas, sustancias, reacciones, propiedades fisicoquímicas, e incluso literatura científica relacionada. Mencionaré ahora las más importantes y las que serán objeto de interés. *PubChem*, una base de datos abierta que sirve información a millones de usuarios en todo el mundo, desde investigadores y estudiantes hasta el público general. Recogen para cada compuesto, información sobre su estructura, representaciones 2D y 3D, identificadores, propiedades químicas y físicas, patentes, avisos de toxicidad, etc. [14]

SciFinder, una herramienta de investigación muy potente que permite explorar las bases de datos de CAS (American Chemical Society) las cuales contienen literatura sobre Química y otras disciplinas afines como Física, Biomedicina, Geología, Ingeniería Química, etc. Incluye referencias bibliográficas y resúmenes de artículos, informes, y libros entre otras cosas. Permite realizar búsquedas por estructura, nombres de sustancias o identificadores, reacciones en la que participa dicha sustancia, artículos y publicaciones que nombren el compuesto en cuestión, e incluso proveedores de compra [15]. Para el uso de esta herramienta es necesario acceder mediante la red de una institución autorizada (en este caso trabajo mediante VPN de

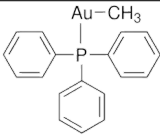
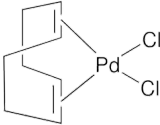
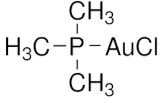
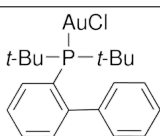
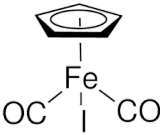
Código SMILES	Representación 2D
<chem>C[Au].c1ccc(cc1)P(c2ccccc2)c3ccccc3</chem>	
<chem>Cl[Pd]Cl.C1CC=CCCC=C1</chem>	
<chem>Cl[Au].CP(C)C</chem>	
<chem>Cl[Au].CC(C)(C)P(c1ccccc1-c2ccccc2)C(C)(C)C</chem>	
<chem>[Fe]I.[C-]#[O+].[C-]#[O+].[CH]1[CH][CH][CH][CH]1</chem>	

Tabla 1.1: Códigos SMILES y sus representaciones visuales según Sigma-Aldrich

la UGR) y seguir los pasos para registrarte ¹. Y *Sigma-Aldrich*, una compañía de ciencia, química y biotecnología que se dedica a la producción y venta de productos químicos, reactivos, equipos y materiales de laboratorio. Ofrece herramientas, servicios, artículos y una gran variedad de productos químicos que se utilizan en investigación, biofarmacéutica, e industria entre otros ámbitos [16]. A través de su página web se enfocan al comercio electrónico pudiendo buscar y comprar productos, compuestos orgánicos e inorgánicos, agentes reactivos, isótopos para síntesis químicas, proteínas, enzimas, etc. De cada producto muestra información relevante como la ficha de datos de seguridad, detalles de las propiedades físicas y químicas así como algunas representaciones lineales del compuesto y la representación del grafo molecular, que es lo interesante en este caso realmente.

Desde la Universidad de Granada, la tutora de este TFG colabora con el grupo de investigación de químicos del ICIQ (Instituto Catalán de Investigación Química) liderado por la profesora Mónica H. Pérez-Temprano. Su foco de investigación gira en torno al entendimiento de transformaciones catalíticas en las que participan compuestos organometálicos, descubriendo y

¹Pasos para el registro en SciFinder https://bibliotecaugr.libguides.com/scifinder_scholar

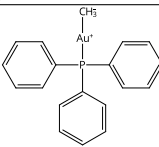
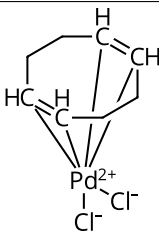
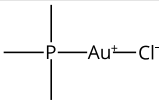
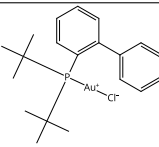
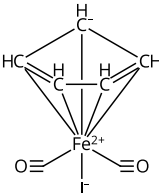
Código SMILES	Representación 2D
<chem>[Au+](CH3-)[P](C=1C=CC=CC1)(C=2C=CC=CC2)C=3C=CC=CC3</chem>	
<chem>[Cl-][Pd+2]123([Cl-])[CH]=4CC[CH]3=[CH]2CC[CH]41</chem>	
<chem>[Cl-][Au+][P](C)(C)C</chem>	
<chem>[Cl-][Au+][P](C=1C=CC=CC1C=2C=CC=CC2)(C(C)(C)C)C(C)(C)C</chem>	
<chem>O#C[Fe+2]1234([I-])(C#O)[CH]=5[CH]4=[CH]3[CH-]2[CH]51</chem>	

Tabla 1.2: Códigos SMILES y sus representaciones visuales según SciFinder

diseñando reacciones más eficientes basadas en catalizadores metálicos. Para más detalle sobre el grupo de investigación y sus ámbitos de trabajo, ver su sitio web [13]. En resumen, intentan desarrollar enfoques más sostenibles para la síntesis de moléculas orgánicas usando la química organometálica. Como tal, necesitan codificar correctamente una molécula de organometálica en pos de trabajar con ella adecuadamente y utilizar todas las herramientas, para, entre otras cosas, poder dibujarla y entenderla mejor.

Uno de los principales problemas que se detectan en este ámbito es la heterogeneidad en las distintas bases de datos para un mismo compuesto o molécula. Para ilustrar esto, presento las tablas 1.1 y 1.2. Ambas tablas comparan las mismas moléculas, mostrando el código SMILES y la representación visual que ofrecen las bases de datos Sigma-Aldrich y SciFinder respectivamente. Vemos diferencias claras en el tratamiento de los ciclos aromáticos, la especificación de las cargas de los átomos y la posición de algunas ramificaciones. Utilizo un subset de 5 moléculas pertenecientes a

la organometálica, seleccionadas desde un conjunto de datos de 30 moléculas considerados de interés por los químicos con los que colabora la tutora (disponible para su consulta en mi GitHub ²). En el Apéndice A, se puede consultar una tabla comparativa con el set de moléculas al completo.

1.2. Objetivos

Por tanto, el objetivo principal de este Trabajo Fin de Grado sería modificar el paquete open-source OpenBabel creando un método para canonizar códigos SMILES, orientado específicamente para compuestos organometálicos. Para ello, se establecen los siguientes subobjetivos:

- Analizar y comparar las cadenas SMILES de distintas bases de datos (p.ej. Sigma-Aldrich, SciFinder) viendo los posibles sinónimos para una misma molécula.
- Determinar un sistema que genere, a partir de cualquier sinónimo SMILES de la misma molécula, un único SMILES canónico.
- Definir otro algoritmo o conjunto de reglas que mejore, aunque sea mínimamente, el sistema de dibujado de las moléculas.

1.3. Estructura de la memoria

esperar a tenerla mas avanzada para completar esto

²<https://github.com/Jesnm01/TFG>

Capítulo 2

Estado del arte y fundamentos teóricos

Quizas sea mejor mover esta seccion justo despues de la introduccion para seguir con la tematica de la motivacion, y ya luego me meto con la gestion y planificacion

Puedo hacer una revision de la literatura existente hasta dia de hoy sobre el tema Usar SCOPUS para esto, con terminos tipo: "SMILESmolecule organometalic" (juntarlos o separarlos segun vea)

Hablar por aqui de la organometalica, representacion de moléculas, Hablar mas extendido de SMILES, SELFIES, e INCHI; cosas de dibujado de moleculas (los paquetes que hay),) No se si meterlo aqui o en otro apartado, el diagrama de clases

la historia de la humanidad está marcada por la búsqueda de materiales que mejoren su calidad de vida, y los metales han sido parte crucial de esos cambios

La materia que forma los seres vivos tiene en su composición sustancias cuya base principal es el carbono. El estudio de estos compuestos constituye una rama de la química llamada química orgánica. La abundancia del carbono en el planeta es relativamente pequeña: aproximadamente un 0,03 %; sin embargo, da lugar a millones de sustancias diferentes, mientras que los compuestos inorgánicos son solo unos pocos miles. ¿Qué hace a este elemento tan especial? Su estructura singular, que le permite formar largas cadenas en las cuales una pequeña variación da lugar a un compuesto distinto al anterior.

Capítulo 3

Gestión y Planificación del proyecto

3.1. Metodología

3.2. Gestión de recursos

3.2.1. Recursos humanos

3.2.2. Recursos materiales

3.2.3. Recursos software

3.3. Gestión de costes

3.3.1. Coste de recursos humanos

Esto quizás dejarlo para el final, cuando tenga la cantidad total de horas trabajadas (aunque podría hacerlo con las "300" horas que se supone que le tengo que dedicar)

3.3.2. Coste hardware

3.3.3. Otros costes

3.3.4. Presupuesto final

3.4. Análisis de riesgos

Capítulo 4

Experimentación

Aquí la idea es ir poniendo las pruebas que vaya haciendo de las moléculas, y lo que vaya descubriendo. Probablemente exponer aquí tb el método o reglas de canonización a las que llegue.

Bibliografía

- [1] William J. Wiswesser. “107 Years of Line-Formula Notations (1861-1968)”. en. En: *Journal of Chemical Documentation* 8.3 (ago. de 1968), págs. 146-150. ISSN: 0021-9576, 1541-5732. DOI: 10.1021/c160030a007. URL: <https://pubs.acs.org/doi/abs/10.1021/c160030a007>.
- [2] Reiner Luckenbach. “The Beilstein Handbook of Organic Chemistry: the first hundred years”. en. En: *Journal of Chemical Information and Computer Sciences* 21.2 (mayo de 1981), págs. 82-83. ISSN: 0095-2338. DOI: 10.1021/ci00030a006. URL: <https://pubs.acs.org/doi/abs/10.1021/ci00030a006>.
- [3] David Weininger. “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules”. en. En: *Journal of Chemical Information and Modeling* 28.1 (feb. de 1988), págs. 31-36. ISSN: 1549-9596. DOI: 10.1021/ci00057a005. URL: <https://pubs.acs.org/doi/abs/10.1021/ci00057a005> (visitado 20-11-2022).
- [4] Johann Gasteiger y Thomas Engel. *Chemoinformatics: A Textbook*. en. John Wiley & Sons, dic. de 2006. ISBN: 978-3-527-60650-4.
- [5] Andrew R. Leach y V. J. Gillet. *An Introduction to Chemoinformatics*. en. Google-Books-ID: 4z7Q87HgBdwC. Springer, sep. de 2007. ISBN: 978-1-4020-6291-9.
- [6] Johann Gasteiger. “The Scope of Chemoinformatics”. en. En: *Handbook of Chemoinformatics*. Ed. por Johann Gasteiger. Weinheim, Germany: Wiley-VCH Verlag GmbH, mayo de 2008, págs. 3-5. ISBN: 978-3-527-61827-9 978-3-527-30680-0. DOI: 10.1002/9783527618279.ch1. URL: <https://onlinelibrary.wiley.com/doi/10.1002/9783527618279.ch1>.
- [7] Nathan Brown. “Chemoinformatics—an introduction for computer scientists”. en. En: *ACM Computing Surveys* 41.2 (feb. de 2009), págs. 1-38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/1459352.1459353. URL: <https://dl.acm.org/doi/10.1145/1459352.1459353>.

- [8] Stephen R. Heller et al. "InChI, the IUPAC International Chemical Identifier". En: *Journal of Cheminformatics* 7.1 (mayo de 2015), pág. 23. ISSN: 1758-2946. DOI: 10.1186/s13321-015-0068-4. URL: <https://doi.org/10.1186/s13321-015-0068-4>.
- [9] Thomas Engel y Johann Gasteiger. *Applied Chemoinformatics. Achievements and Future Opportunities*. en. Wiley-VCH, 2018. ISBN: 978-3-527-34201-3.
- [10] Mario Krenn et al. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation". en. En: *Machine Learning: Science and Technology* 1.4 (dic. de 2020), pág. 045024. ISSN: 2632-2153. DOI: 10.1088/2632-2153/aba947. URL: <https://iopscience.iop.org/article/10.1088/2632-2153/aba947>.
- [11] Mario Krenn et al. "SELFIES and the future of molecular string representations". en. En: *Patterns* 3.10 (oct. de 2022). arXiv:2204.00056 [physics], pág. 100588. ISSN: 26663899. DOI: 10.1016/j.patter.2022.100588. URL: <http://arxiv.org/abs/2204.00056> (visitado 20-11-2022).
- [12] Alston Lo et al. *Recent advances in the Self-Referencing Embedding Strings (SELFIES) library*. arXiv:2302.03620 [physics]. Feb. de 2023. DOI: 10.48550/arXiv.2302.03620. URL: <http://arxiv.org/abs/2302.03620> (visitado 26-02-2023).
- [13] ICIQ. Prof. Mónica H. Pérez-Temprano, Research Group. en. URL: https://www.iciq.org/research/research_group/dr-monica-h-perez-temprano/section/research_overview/ (visitado 10-03-2023).
- [14] PubChem. *PubChem*. en. URL: <https://pubchem.ncbi.nlm.nih.gov> (visitado 01-02-2023).
- [15] SciFinder. *SciFinder-help*. en. URL: https://scifinder-n.cas.org/help/#t=Working_with_Search_Results%5C%2FAll_Answer_Types_screen.htm (visitado 01-02-2023).
- [16] SigmaAldrich. *SigmaAldrich - About us*. es. URL: <https://www.sigmaaldrich.com/ES/es/life-science/about-us/expertise> (visitado 01-02-2023).

Apéndice A

Tabla comparativa del set completo de moléculas

Tabla A.1: Tabla de índices con los nombres de las moléculas de la Tabla A.2

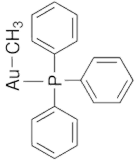
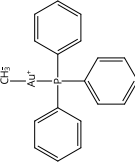
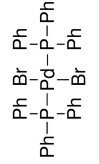
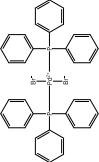
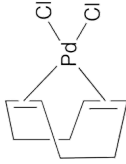
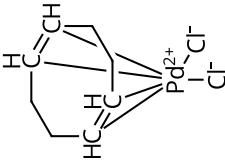
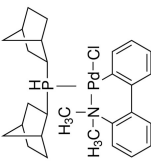
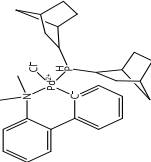
Id	Nombre del compuesto
1	Methyl(triphenylphosphine)gold(I)
2	trans-Dibromobis(triphenylphosphine)palladium(II)
3	Dichloro(1,5-cyclooctadiene)palladium(II)
4	SK-CC 01A
5	Bis[μ-chloro[5-hydroxy-2-[1-(hydroxyimino)ethyl]phenyl]palladium]
6	(SP-4-3)-(3,5-Dichloro-2,4,6-trifluorophenyl)iodobis (triphenylarsine)palladium
7	Palladium,(2,2'-bipyridine- <i>k</i> N1, <i>k</i> N1')[(2,2-dimethyl-1,2-ethanediyl)-1,2-phenylene]fluoro(4-methylbenzenesulfonamidato- <i>k</i> N)-, (OC-6-35)
8	Dibromo(1,2-dimethoxyethane)nickel(II)
9	Bis(triphenylphosphine)ruthenium(II) dicarbonyl chloride
10	Chloro(trimethylphosphine)gold(I)
11	Chloro[tris(para-trifluoromethylphenyl)phosphine]gold(I)
12	Chloro(dimethylsulfide)gold(I)

Continúa en la siguiente página

Tabla A.1 – Continuación de la página anterior

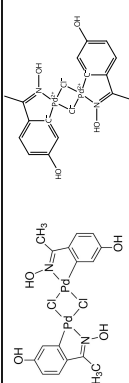
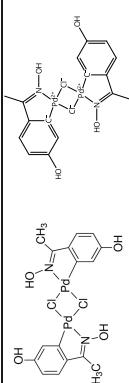
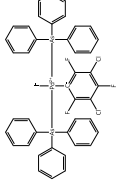
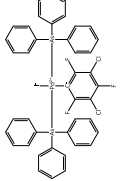
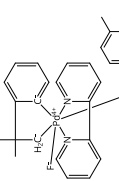
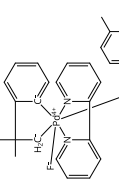
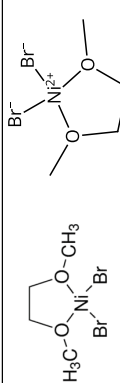
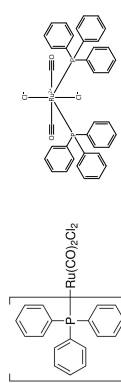
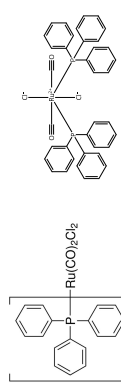
Id	Nombre del compuesto
13	Chloro(methyldiphenylphosphine)gold(I)
14	Chloro[diphenyl(o-tolyl)phosphine]gold(I)
15	Chloro[(1,1'-biphenyl-2-yl)di-tert-butylphosphine]gold(I)
16	(Acetonitrile)[(2-biphenyl)di-tert-butylphosphine] gold(I)hexafluoroantimonate
17	Chloro[2-di-tert-butyl(2',4',6'-triisopropylbiphenyl)phosphine] gold(I)
18	Chloro[2-dicyclohexyl(2',4',6'-trisopropylbiphenyl)phosphine]gold(I)
19	BisPhePhos XD gold(I) chloride
20	Chloro[di(1-adamantyl)-2-dimethylaminophenylphosphine]gold(I)
21	Dichloro(DPPE)digold(I)
22	Dichloro[(±)-BINAP]digold(I)
23	Bis(chlorogold(I)) [1,1'-bis(diphenylphosphino)ferrocene]
24	[(IMes)AuCl]
25	[(IPr)AuCl]
26	IPrAuNTf ₂
27	DPPF
28	Dicarbonylcyclopentadienyliodoiron(II)
29	(OC-6-11')-Bis[2,6-di(2-pyridinyl- <i>k</i> N)phenyl- <i>k</i> C]iron

Tabla A.2: Tabla extendida para el set de datos de 30 moléculas. Contiene la cadena SMILES extraída de Sigma-Aldrich (SA), la cadena SMILES extraída de SciFinder (SF), y las imágenes de las respectivas bases de datos (SA y SF)

Id	SMILES SA	SMILES SF	Imagen SA	Imagen SF
1	<chem>C[Au].c1ccc(cc1)P(c2ccccc2)c3ccccc3</chem>	<chem>[Au+](([CH3-])[P](C=1C=CC=CC1)(C=2C=CC=CC2)C=3C=CC=CC3</chem>		
2	<chem>Br[Pd]Br.c1ccc(cc1)P(c2ccccc2)c3ccccc3.c4ccc(cc4)P(c5ccccc5)c6ccccc6</chem>	<chem>[Br-][Pd+2]([Br-])([P](C=1C=CC=CC1)(C=2C=CC=CC2)C=3C=CC=CC3)[P](C=4C=CC=CC4)(C=5C=CC=CC5)C=6C=CC=CC6</chem>		
3	<chem>Cl[Pd]Cl.C1CC=CCCC=C1</chem>	<chem>[Cl-][Pd+2]123([Cl-])[CH]=4CC[CH]3=[CH]2CC[CH]41</chem>		
4	<chem>C1C[C@@H]2C[C@H]1CC2PC3C[C@@H]4CC[C@H]3C4.CN(C)c5ccccc5c6ccccc6[Pd]Cl</chem>	<chem>[Cl-][Pd+2]1([C-]=2C=CC=CC2C=3C=CC=CC3[N]1(C)C)[PH](C4CC5CCC4C5)C6CC7CCC6C7</chem>		

Continúa en la siguiente página

Tabla A.2 – Continuación de la página anterior

Id	SMILES SA	SMILES SF	Imagen SA	Imagen SF
5	<chem>C\C(=N/O)c1ccc(O)cc1[Pd]Cl</chem> <chem>.C\C(=N/O)c2ccc(O)cc2[Pd]Cl</chem>	<chem>OC=1C=CC=2C(=[N](O)[Pd+2]3</chem> <chem>([Cl-])[Pd+2]4([Cl-]3)[C-]=5</chem> <chem>C=C(O)C=CC5C(=[N]4O)C)[C-]2C1)C</chem>		
6	No se encontró el compuesto en Sigma-Aldrich	<chem>FC=1C(Cl)=C(F)[C-](=C(F)ClCl)[Pd+2]</chem> <chem>([F-])([As])(C=2C=CC=CC2)(C=3C=CC=C</chem> <chem>C3)C=4C=CC=CC4)[As](C=5C=CC=CC5)</chem> <chem>(C=6C=CC=CC6)C=7C=CC=CC7</chem>		
7	No se encontró el compuesto en Sigma-Aldrich	<chem>O=S(=O)([NH-][Pd+4]12([F-])([C-</chem> <chem>]=3C=CC=CC3C(C)(C)[CH2-]1)</chem> <chem>[N]=4C=CC=CC4C=5C=CC=C[N]</chem> <chem>52)C6=CC=C(C=C6)C</chem>		
8	<chem>Br[Ni]Br.COCCOC</chem>	<chem>[Br-][Ni+2]1([Br-])O(C)CCO1C</chem>		
9	<chem>Cl[Ru](Cl)(C#[O])(C#[O])([PH](c1ccc</chem> <chem>cc1)(c2ccccc2)c3ccccc3)[PH](c4ccc</chem> <chem>cc4)(c5ccccc5)c6ccccc6</chem>	<chem>O#C[Ru+2]([Cl-])([Cl-])([Cl-</chem> <chem>])(C#O)([P](C=1C=CC=CC1)</chem> <chem>(C=2C=CC=CC2)C=3C=CC=CC3)[P]</chem> <chem>(C=4C=CC=CC4)(C=5C=CC=CC5)</chem> <chem>C=6C=CC=CC6</chem>		

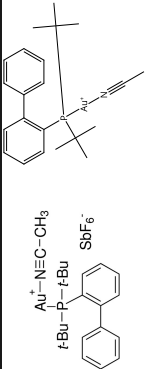
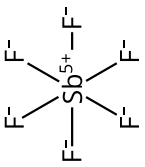
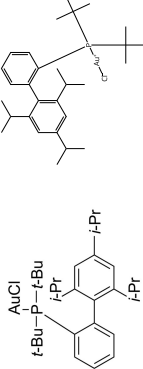
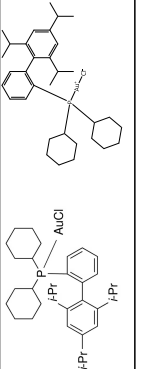
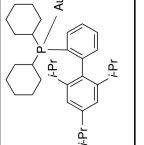
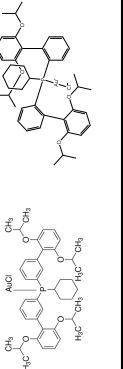
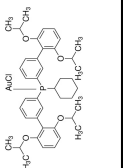
Continúa en la siguiente página

Tabla A.2 – Continuación de la página anterior

Id	SMILES SA	SMILES SF	Imagen SA	Imagen SF
10	<chem>Cl[Au].CP(C)C</chem>	<chem>[Cl-][Au+][P](C)(C)C</chem>		
11	<chem>Cl[Au].FC(F)(F)c1ccc(cc1)P(c2ccc(cc2)C(F)(F)F)c3ccc(cc3)C(F)(F)F</chem>	<chem>FC(F)(F)C1=CC=C(C(C=C1)[P]([Au+][Cl-])(C2=CC=C(C(C=C2)C(F)(F)F)C3=CC=C(C(C=C3)C(F)(F)F)F</chem>		
12	<chem>Cl[Au].CSC</chem>	<chem>[Cl-][Au+][S](C)C</chem>		
13	<chem>Cl[Au].CP(c1cccc1)c2cccc2</chem>	<chem>[Cl-][Au+][P](C=1C=CC=CC1)(C=2C=CC=CC2)C</chem>		
14	<chem>Cl[Au].Cc1cccc1P(c2cccc2)c3cccc3</chem>	<chem>[Cl-][Au+][P](C=1C=CC=CC1)(C=2C=CC=CC2)C=3C=CC=CC3C</chem>		
15	<chem>Cl[Au].CC(C)(C)P(c1cccc1-c2cccc2)C(C)(C)C</chem>	<chem>[Cl-][Au+][P](C=1C=CC=CC1C=2C=CC=CC2)(C(C)(C)C)C(C)(C)C</chem>		

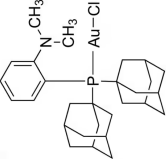
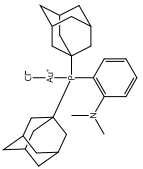
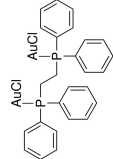
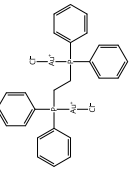
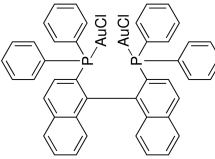
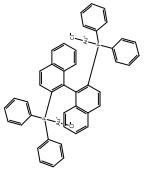
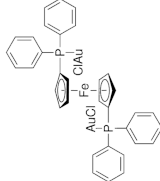
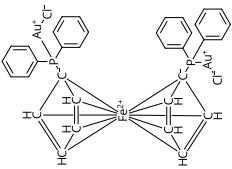
Continúa en la siguiente página

Tabla A.2 – Continuación de la página anterior

Id	SMILES SA	SMILES SF	Imagen SA	Imagen SF
16	<chem>[Au+].CC#N.F[Sb-](F)(F)(F)F. CC(C)(C)P(c1cccc1- c2cccc2)C(C)(C)C</chem>	<chem>[F-][Sb+5]([F-])([F-])([F-])([F-])([F-])[F-]. C([#N])[Au+][P](C=1C=CC=CC1C= 2C=CC=CC2)(C(C)(C)C(C)C(C)C)C</chem>		
17	<chem>CC(C)c1cc(C(C)C)c(c(c1)C(C)C)-c2 cccc2[PH]([Au]Cl)(C(C)(C)C)C(C)C</chem>	<chem>[Cl-][Au+][P](C=1C=CC=CC1C=2C (=CC(=CC2C(C)C)C(C)C)C(C)C)(C(C) (C)C)C(C)(C)C</chem>		
18	<chem>Cl[Au].CC(C)c1cc(C(C)C)c(c(c1)C(C) C)-c2cccc2P(C3CCCCC3)C4CCCCC4</chem>	<chem>[Cl-][Au+][P](C=1C=CC=CC1C=2C (=CC(=CC2C(C)C)C(C)C)C(C)C)(C3C CCCC3)C4CCCCC4</chem>		
19	<chem>CC(C)OC(C=CC=C1OC(C)C)=C1C2 =CC(P(C3CCCCC3)O4=CC(O5=C(O C(C)C)C=CC=C5OC(C)C)=CC=C4) =CC=C2.[Au]Cl</chem>	<chem>[Cl-][Au+][P](C=1C=CC=CC1C2=C(OC(C) C)C=CC=C2OC(C)C)(C=3C=CC=CC3C4 =C(OC(C)C)C=CC=C4OC(C)C)C5CCCCC5</chem>		

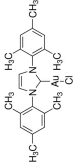
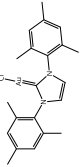
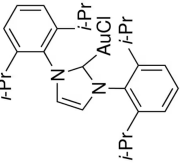
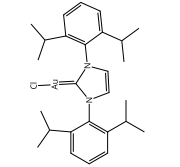
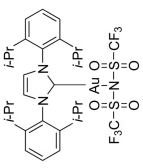
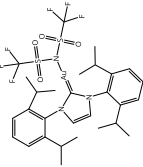
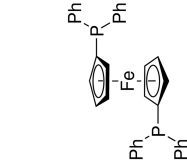
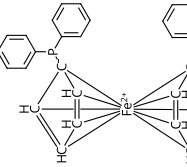
Continúa en la siguiente página

Tabla A.2 – Continuación de la página anterior

Id	SMILES SA	SMILES SF	Imagen SA	Imagen SF
20	<chem>Cl[Au].CN(C)c1cccc1P(C23C C4CC(CC(C4)C2)C3)C56CC7CC(C C(C7)C5)C6</chem>	<chem>[Cl-].[Au+].[P](C=1C=CC=CC1N (C)C)(C23CC4CC(CC(C4)C2)C3) C56CC7CC(CC(C7)C5)C6</chem>		
21	<chem>Cl[Au].Cl[Au].C((CP(c1cccc1) c2cccc2)P(c3cccc3)c4cccc4</chem>	<chem>[Cl-].[Au+].[P](C=1C=CC=CC1) (C=2C=CC=CC2)CC[P]([Au+][Cl-]) (C=3C=CC=CC3)C=4C=CC=CC4</chem>		
22	<chem>Cl[Au].Cl[Au].P(C1=CC=CC=C1)(C2 =C(C3=C(C=CC4=C3C=CC=C4)P (C5=CC=CC=C5)C6=CC=CC=C6)C7 =C(C=CC=C7)C8=CC=CC=C8</chem>	<chem>[Cl-].[Au+].[P](C=1C=CC=CC1)(C=2C =CC=CC2)C3=CC=C4C=CC=CC4=C3 C=5C=6C=CC=CC6C=CC5[P]([Au+][Cl-])(C=7C=CC=CC7)C=8C=CC=CC8</chem>		
23	<chem>[Fe].Cl[Au].Cl[Au].[CH]1[CH] [CH][C]([CH]1)P(c2cccc2)c3 ccccc3.[CH]4[CH][CH][C]([CH]4) P(c5cccc5)c6cccc6</chem>	<chem>[Cl-].[Au+].[P](C=1C=CC=CC1)(C=2C=CC =CC2)[C-]34[CH]5=[CH]6[CH]7=[CH]3[Fe+2] 6789%10%1154[CH]=%12[CH]%11=[CH]%10 [C-]9([CH]%128)[P]([Au+][Cl-])(C=%13 C=CC=CC%13)C=%14C=CC=CC%14</chem>		

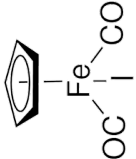
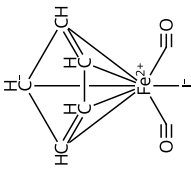
Continúa en la siguiente página

Tabla A.2 – Continuación de la página anterior

Id	SMILES SA	SMILES SF	Imagen SA	Imagen SF
24	<chem>Cl[Au].Cc1cc(C)c(N2[C]N(C=C2)c3c(C)cc(C)cc3C)c(C)c1</chem>	<chem>Cl[Au]=C1N(C=CN1C=2C(=CC(=CC2C(C)C)C=3C(=CC(=CC3C)C)C</chem>		
25	<chem>CC(C)c1cccc(C(C)C)c1N2C=CN(C2[Au]Cl)c3c(ccc3C(C)C)C(C)C</chem>	<chem>Cl[Au]=C1N(C=CN1C=2C(=CC=CC2C(C)C)C(C)C=3C(=CC=CC3C(C)C)C(C)C</chem>		
26	<chem>CC(C)c1cccc(C(C)C)c1N2C=CN(C2=[Au]N(S(=O)(=O)C(F)(F)F)S(=O)(=O)C(F)(F)F)c3c(ccc3C(C)C)C(C)C</chem>	<chem>O=S(=O)(N([Au]=C1N(C=CN1C=2C(=CC=CC2C(C)C)C(C)C=3C(=CC=CC3C(C)C)C(C)C=4C(=CC(=CC4C(F)F)C(F)F)C(F)F)F</chem>		
27	<chem>[Fe].[CH]1[CH][CH][C]([CH]1)P(c2cccc2)c3cccc3.[CH]4[CH][CH][C]([CH]4)P(c5cccc5)c6cccc6</chem>	<chem>C=1C=CC(=CC1)P(C=2C=CC=CC2)[C-]34[CH]5=[CH]6[CH]7=[CH]3[Fe+2]6789%10%1154[CH]=%12[CH]%11=[CH]%10[C-]9[P(C=%13C=CC=CC%13)C=%14C=CC=CC%14][CH]%128</chem>		

Continúa en la siguiente página

Tabla A.2 – Continuación de la página anterior

Id	SMILES SA	SMILES SF	Imagen SA	Imagen SF
28	<chem>[Fe]I.[C-]#[O+].[C-]#[O+].[CH]1[CH][CH][CH][CH]1</chem>	<chem>O#C[Fe+2]1234([I-])(C#O)[CH]=5[CH]4=[CH]3[CH-]2[CH]51</chem>		
29	No se encontró el compuesto en Sigma-Aldrich	<chem>C=1C=C[N]2=C(C1)C3=CC=CC=4C=5C=CC=C[N]5[Fe+2]672([C-]34)[C-]=8C(=CC=CC8C=9C=CC=C[N]96)C=%10C=CC=C[N] %107</chem>		