

QUÍMICA DE PERFUMES, TRADICIÓN E INNOVACIÓN: CÓMO MACHINE LEARNING PUEDE AYUDAR A MEJORAR LOS PROCESOS DE CREACIÓN DE NUEVOS AROMAS

Como menciona Krenn (2019), los modelos generativos profundos han ganado una atención considerable en años recientes con resultados notables en generación de imágenes, ciencia de materiales, y una apertura de un nuevo diseño computacional para nuevos medicamentos y materiales. Estos modelos son, generalmente, estructuras complejas expresadas como grafos en restricciones semánticas del dominio al que sean aplicadas; y se adhieren al Machine Learning como preparación para la evaluación de valores posteriores. El área química se ha encontrado con una falta de beneficios desde las nuevas tecnologías por su carácter observatorio y no predictivo, fenómeno que se debate éticamente dificulta su adaptación y aplicación, impidiendo nuevas oportunidades laborales, como menciona Brynjolfsson (2019). Es necesario entonces, para la perfumería, ver su estado de evolución tecnológica hacia la optimización de sus procesos. Dice Paschek et al. (2017) que la digitalización actual es imparable y cambia la situación del mercado en muchos sectores empresariales, citando a August Wilhelm Scheer, Director Gerente del Grupo Scheer, en el 2015: “todo lo que se pueda digitalizar se digitalizará. Cada proceso, producto y modelo de negocio se convierte en digital, por lo que la revolución radica en los procesos ”.

Hasta hace poco, la química de perfumería se ha centrado en invertir para asegurar la calidad de sus productos mediante revisión de evaluaciones y un Sistema de Administración Integrado, como menciona Razmochaeva, Semenov y Bezrukov (2019), que se divide en un Sistema de Administración para la Calidad (QMS), un Sistema de Administración Ambiental (EMS), y un Sistema de Administración de Salud y Seguridad (SHMS), basados en los estándares internacionales ISO 9001: 2008, ISO 14001: 2004, y OHSAS 18001: 2007. Estos sistemas están documentados para el manejo de la calidad, seguridad, salud, objetivos relevantes, procedimientos, instrucciones de trabajo y récords específicos para las políticas específicas de cada compañía y que se aplica a, por ejemplo, auditar proveedores al mismo tiempo que internamente se montan criterios de aceptación, almacenamiento, envío, reempaque y distribución de bienes.

Una parte fundamental para asegurar dicha calidad de productos es una buena obtención de materia prima con la que trabajar, cuyos métodos típicos para la obtención y combinación de fragancias están basadas en diferentes técnicas físicas y químicas, como la destilación y el enfleurage. Entrando en detalle sobre la destilación, Vogelpohl (2015) define al principio de la misma como el ocasionar la evaporación de una mezcla en ebullición para separar una mezcla líquida en fracciones con composiciones diferentes de la mezcla líquida e incluso en sus componente (Vogelpohl, 2015). La Figura 1 muestra una configuración para llevar a cabo una destilación simple, que consiste en una olla destilada calentada, un condensador para licuar el vapor producido en la olla destilada y un receptor para recoger el destilado.

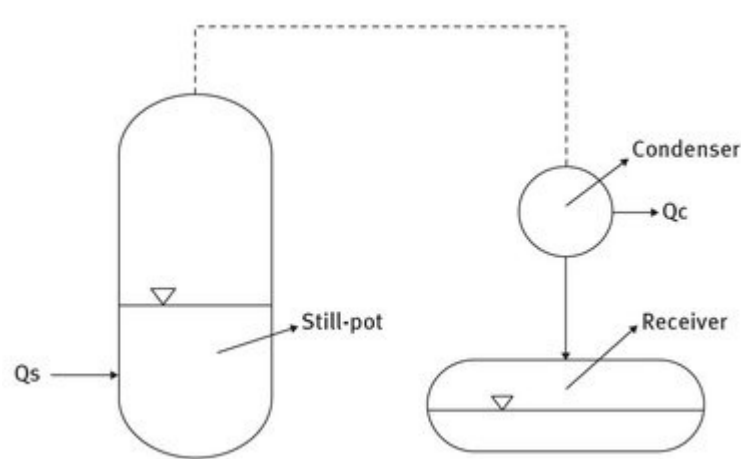


Figura 1. Representación lógica de la configuración de un destilador. Reimpreso de Distillation, A.Vogelpohl, 2015.

Si se agrega calor continuamente a la olla fija, parte del líquido en la olla fija se vaporiza y suponiendo que los componentes de la mezcla líquida tengan una presión de vapor diferente, el vapor que sale de la olla fija se enriquecerá en los componentes. con una presión de vapor más alta que resulta en un destilado en el receptor diferente de la mezcla líquida en el recipiente inmóvil (Vogelpohl, 2015); luego, las materias primas extraídas utilizando la destilación y otros métodos también tradicionales y otros modernos (como el SOFTACT) son mezcladas meticulosamente para generar las fragancias finales por los perfumistas. Para las mezclas de olores, se informan muchos modelos como el modelo ERM, el modelo vectorial, el modelo U, el modelo de aditividad y sus versiones extendidas. Estos modelos proporcionan ideas valiosas y orientación para nuestra comprensión de la interacción del olor, pero su precisión de predicción y alcance aplicable a menudo son limitados ante, por ejemplo, la dificultad de predecir de la intensidad del olor de diferentes mezclas (Yan et al., 2020).

Ahora, con la Revolución Industrial 4.0, la química hacia sus campos de investigación ha hecho avances como el reconocimiento y la predicción de la estructura química en tiempo real. Keyrouz, Tauk y Feghali (2018), por ejemplo, diseñaron un sistema que despliega dos redes bayesianas que, con una unidad de procesamiento morfológico para eliminar el ruido, regulariza formas y acentuar contornos que superan a las técnicas convencionales basadas en Análisis de Componentes Principales (PCA) y Análisis Lineal Discriminante (LDA). También se están empleando Máquinas de Soporte de Vectores (SVM) gracias a su capacidad de clasificar datos lineales y no lineales. La figura 2 presenta una gráfica típica de cómo se emplean los datos para predecir la una línea de salida, llamada hiperplano (hyperplane). El principio de las máquinas de soporte de vectores es mapear cada elemento de datos en un espacio de características n-dimensional donde n es el número de características para luego identificar el hiperplano que separa los elementos de datos en dos clases mientras maximiza la distancia marginal para ambas clases y minimiza los errores de clasificación (Uddin et al., 2019), método que ha tenido un éxito significativo en numerosas tareas de aprendizaje del mundo real para procesos químicos como el modelo Smooth Transition Regression (STR) evaluadas por Yu et al. (2002), puesto que se emplean seleccionando los primeros sensores de semillas según el criterio de margen máximo entre las diferentes clases de olores. Estos sensores identificados se utilizan posteriormente como candidato inicial en el algoritmo de búsqueda. A partir de los resultados experimentales en el conjunto de datos de refrescos, el número de sensores seleccionados no solo se reduce significativamente, sino que también se incrementa el rendimiento de clasificación (Phaisangittisagul, 2010).

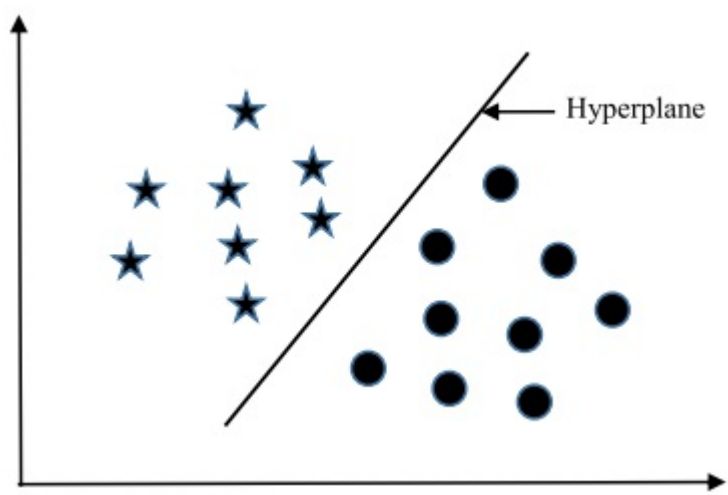


Figura 2. Ilustración simplificada de cómo funciona la máquina de vectores de soporte. Recuperado de Comparing different supervised machine learning algorithms for disease prediction.

Uddin, S., Khan, A., Hossain, M., & Moni, M. (2019).

Yan et al. (2020) utilizó un algoritmo de regresión de vectores de soporte (modelo que se ha venido practicando frecuentemente en el área perfumista) para establecer modelos de predicción de intensidad de olor para ésteres binarios, aldehídos y mezclas de hidrocarburos aromáticos, todos mencionados en la tabla 1, y lo adaptó para generar datos de olor adicionales al predecir la intensidad del olor de muestras más simuladas con diversas relaciones de mezcla y niveles de concentración.

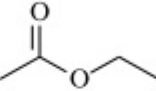
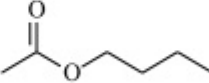
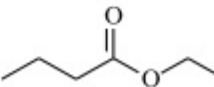
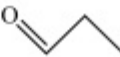
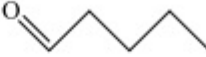



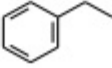
Tabla 1. Muestras de entrenamiento de inserción para el set de entrenamiento con base al algoritmo SVR.

Readaptado de Visual analysis of odor interaction based on support vector regression method. Yan et al. (2020)

Orden	Odorante	Estructura Química
(Abreviación)		

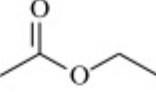
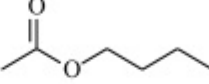
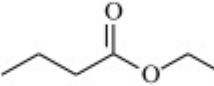





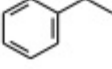
1

acetato de etilo
(EA)

Order	Odorant (Abbreviation)	CAS#	Chemical Structure
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

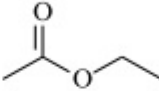
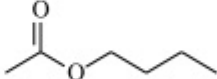
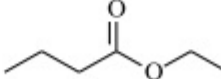
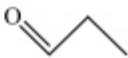
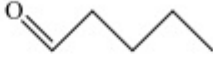



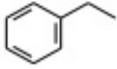
2

acetato de
butilo (BA)

Order	Odorant (Abbreviation)	CAS#	Chemical Structure
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

3

butirato de etilo
(EB)

Order	Odorant (Abbreviation)	CAS#	Chemical Structure
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

4

propionaldehído
(PA)

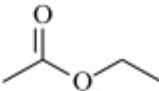
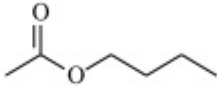
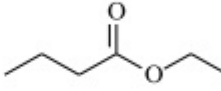
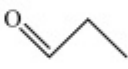




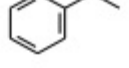
Order	Odorant (Abbreviation)	CAS#	Chemical Structure
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

Order	Odorant (Abbreviation)	CAS#	Chemical Structure
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

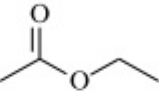
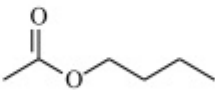
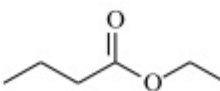
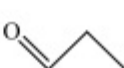




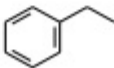
5

n-valeraldehído (VA)

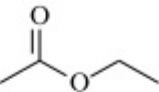
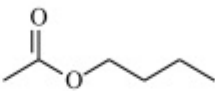
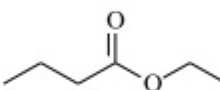





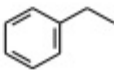
n-heptaldehydo
(HEP)

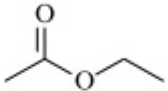
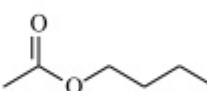
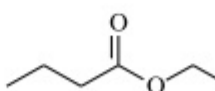
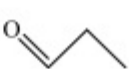
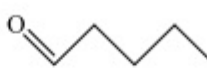



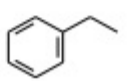
Order	Odorant (Abbreviation)	CAS#	Chemical Struct
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

7 Benceno (B)

Order	Odorant (Abbreviation)	CAS#	Chemical Structure
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

8 Tolueno (T)

Order	Odorant (Abbreviation)	CAS#	Chemical Structure
1	ethyl acetate (EA)	141-78-6	
2	butyl acetate (BA)	123-86-4	
3	ethyl butyrate (EB)	105-54-4	
4	propionaldehyde (PA)	123-38-6	
5	<i>n</i> -valeraldehyde (VA)	110-62-3	
6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
7	benzene (B)	71-43-2	
8	toluene (T)	108-88-3	
9	Ethylbenzene (E)	100-41-4	

		Order	Odorant (Abbreviation)	CAS#	Chemical Structure
9	Etilbencina (E)	1	ethyl acetate (EA)	141-78-6	
		2	butyl acetate (BA)	123-86-4	
		3	ethyl butyrate (EB)	105-54-4	
		4	propionaldehyde (PA)	123-38-6	
		5	<i>n</i> -valeraldehyde (VA)	110-62-3	
		6	<i>n</i> -heptaldehyde (HEP)	117-71-7	
		7	benzene (B)	71-43-2	
		8	toluene (T)	108-88-3	
		9	Ethylbenzene (E)	100-41-4	

Los resultados del experimento de Yan et al. (2020), para la mayoría de las muestras de entrenamiento y muestras de prueba, hicieron con éxito predicciones perfectas, mostrando así la viabilidad y la buena capacidad de ajuste del algoritmo SVR en el análisis de datos de olores regulares. La figura 3 enumera el coeficiente de determinación (R2) y el error absoluto medio (MAE) entre la intensidad del olor olfativo medido y la intensidad del olor predicha por SVR de cada mezcla de olor individualmente.

Es un aspecto curioso denotar que, a diferencia de las otras mezclas, los valores de R2 de la mezcla T + E fueron menores, menciona Yan et al. (2020) que probablemente fue causado por una precisión relativamente pobre de los resultados olfativos medidos porque el algoritmo SVR es muy sensible al ruido en los datos de entrenamiento.

Mixture	R^2		MAE	
	Training Set	Test Set	Training Set	Test Set
EA+BA	0.97	0.95	0.15	0.26
BA+EB	0.96	0.85	0.15	0.33
EA+EB	0.87	0.87	0.17	0.25
PA+VA	0.95	0.78	0.14	0.25
PA+HEP	0.96	0.94	0.17	0.23
VA+HEP	0.97	0.87	0.15	0.31
B+T	0.87	0.81	0.33	0.43
T+E	0.78	0.68	0.31	0.40
B+E	0.98	0.94	0.09	0.27

Figura 3. Resultados de la predictibilidad del modelo entrenado con el algoritmo SVR. Adaptado de Visual Analysis of Odor Interaction Based on Support Vector Regression Method. Sensors, Yan, L., Wu, C., & Liu, J., 2020, 20(6), 1707.

Sin embargo, existen otros tipos de algoritmos utilizados en Machine Learning, como la regresión logística. Como menciona Uddin et al. (2019), empezó como una extensión de la regresión ordinaria y está diseñada para modelar sólo una variable que generalmente representa la ocurrencia o no ocurrencia de un evento para ayudar a encontrar la probabilidad de que una nueva instancia pertenezca a una determinada clase. Como probabilidad que es, sus valores se encuentran entre 0 y 1. Adicionalmente, el modelo puede generalizarse para modelar una variable categórica con más de dos valores. Esta versión se conoce como la regresión logística multinomial.

También se han utilizado redes neuronales, definidos por Uddin et al. (2019) como un conjunto de algoritmos de aprendizaje donde las neuronas están conectadas entre sí, representando un grupo interconectado de nodos. La salida de un nodo pasa como entrada a otro nodo para su posterior procesamiento de acuerdo con la interconexión. Los nodos y los bordes tienen pesos que permiten ajustar la intensidad de la señal de comunicación, debilitándose o amplificándose basado en el entrenamiento y la posterior adaptación de las matrices, los nodos y los pesos de los bordes. Las flechas conectan la salida de nodos de una capa a la entrada de nodos de otra capa.

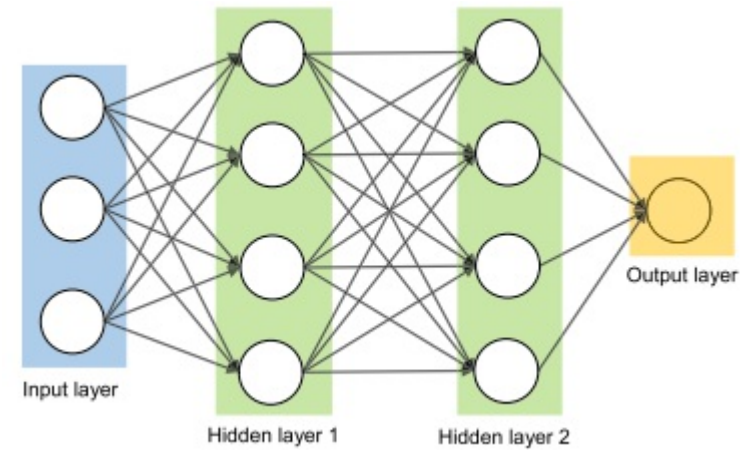


Figura 5. Una ilustración de la estructura de la red neuronal artificial con dos capas ocultas. Recuperado de Comparing different supervised machine learning algorithms for disease prediction. Uddin, S. et al., (2019). BMC Medical Informatics And Decision Making, 19(1).

Los algoritmos ya mencionados, y otros resumidos en la tabla 2, han estado expandiéndose hacia varias aplicaciones dentro de diferentes áreas también relacionadas directamente con la química de una u otra manera, como múltiples investigaciones de modelamiento para la investigación de síntomas clave, reacciones químicas, comportamientos fisiológicos, y predicción de mutación de enfermedades.

Tabla 2. Comparación de tipos de algoritmos utilizados en Machine Learning.

Algoritmo	Contexto	Ventajas	Desafíos
-----------	----------	----------	----------

Regresión logística	Para la clasificación supervisada.	Valor probabilístico	Encuentra la probabilidad de que una instancia pertenezca a una clase.	Se debe asignar un umbral para diferenciar dos clases.
Máquina de Soporte de Vectores (SVM)	Para clasificar datos lineales y no lineales.	Hiperplano que maximiza la separación entre las clases.	Cada punto de datos se traza como un punto en un espacio definido	Se debe encontrar el hiperplano que diferencia las dos clases por el margen máximo.
Clasificador bayesiano ingenuo	Basada en el teorema de Bayes. Supone que una característica particular en una clase no se relaciona directamente con otra.	Combinación de dos fuentes de información. Posterior probabilidad posterior de que una fuente pertenezca a la otra.	Fácil de implementar, no requiere muchos datos de entrenamiento. Rápido. Usado para hacer predicciones en tiempo real.	Si no hay ocurrencias entre una clase y un atributo, la probabilidad estimada será 0, afectando la multiplicación de probabilidades.

Por ejemplo, Stark et al. (2019) desarrolló modelos de aprendizaje automático que utilizaron datos de salud personales para predecir el riesgo de cáncer de seno a cinco años; utilizando entradas del modelo Gail y modelos que usan tanto entradas del modelo Gail como datos de salud personales adicionales relevantes para el riesgo de cáncer de mama. Los resultados del modelo aplicado, según los autores, podrían usarse como herramientas de estratificación de riesgos no invasivas y rentables para aumentar la detección y prevención tempranas del cáncer de seno. También se han evaluado modelos de aprendizaje automático entrenados para predecir la mortalidad con un estudio de caso de asma grave, aunque los experimentos de Stark et al. (2019) para seleccionar menos características de entrada basadas en una puntuación discriminativa mostraron una precisión baja a moderada para descubrir trillizos clínicamente significativos. Estos resultados indican que, si bien indican que la puntuación discriminativa por sí sola no puede reemplazar la entrada clínica, el uso de características filtradas puede reducir la complejidad del modelo y tener poco impacto en el rendimiento de un modelo de Machine Learning.

Se ha revisado entonces acerca de diferentes técnicas y metodologías tanto tradicionales como emergentes para la creación de fragancias, las cumbres y retos que la industria enfrenta en este momento y los aportes que, en general, se han brindado hacia su innovación. SOFTACT es, por ejemplo, apenas y un método nuevo para extracción de fragancias, y se espera ver que la Revolución Tecnológica 4.0 siga impactando en este y otros campos del área química. Con relación al proyecto de investigación, los algoritmos aquí presentados presentan indicios fuertes hacia cómo podrían ser de ayuda hacia la industria química, concatenando las capacidades actuales de ésta área de Data Science y los desafíos a los que se encuentra la perfumería y, más aún, bases sobre las cuales se desarrollará para la investigación de nuevas fronteras.

Referencias

- Berman, F., Stodden, V., Szalay, A., Rutenbar, R., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Martonosi, M. and Raghavan, P. (2018). Realizing the potential of data science. *Communications of the ACM*, 61(4), pp.67-72.
- Berinato, S. (2019). Data Science and the Art of Persuasion. *Harvard Business Review*, 97(1), 126–137.
- Machine Learning Search Term. (2020). Recuperado 12 Febrero 2020, de <https://trends.google.com/trends/explore?date=2010-01-01%202020-01-01&q=machine%20learning>
- Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, 358(6370), 1530-1534. <https://doi.org/10.1126/science.aap8062>
- Nicely, J., Duncan, B., Hanisco, T., Wolfe, G., Salawitch, R., & Deushi, M. et al. (2019). A Machine Learning Examination of Hydroxyl Radical Differences Among Model Simulations for CCMI-1. <https://doi.org/10.5194/acp-2019-772>
- Schütt, K., Gastegger, M., Tkatchenko, A., Müller, K., & Maurer, R. (2019). Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10(1). <https://doi.org/10.1038/s41467-019-12875-2>
- SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. (2020), 1. <https://doi.org/arXiv:1905.13741>
- Paschek, D., Luminosu, C., & Draghici, A. (2017). Automated business process management – in times of digital transformation using machine learning or artificial intelligence. *MATEC Web Of Conferences*, 121, 04007. <https://doi.org/10.1051/mateconf/201712104007>

- Phaisangittisagul, E. (2010). Improving Sensor Subset Selection of Machine Olfaction Using Multi-class SVM. *2010 Third International Conference On Knowledge Discovery And Data Mining*. <https://doi.org/10.1109/wkdd.2010.39>
- Ruiz, F., Agell, N., Angulo, C., & Sánchez, M. (2018). A learning system for adjustment processes based on human sensory perceptions. *Cognitive Systems Research*, 52, 58-66. <https://doi.org/10.1016/j.cogsys.2018.06.011>
- Vogelpohl, A. (2015). *Distillation* (1st ed.). Walter de Gruyter GmbH.
- Yan, L., Wu, C., & Liu, J. (2020). Visual Analysis of Odor Interaction Based on Support Vector Regression Method. *Sensors*, 20(6), 1707. <https://doi.org/10.3390/s20061707>
- Uddin, S., Khan, A., Hossain, M., & Moni, M. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics And Decision Making*, 19(1). <https://doi.org/10.1186/s12911-019-1004-8>
- Stark, G., Hart, G., Nartowt, B., & Deng, J. (2019). Predicting breast cancer risk using personal health data and machine learning models. *PLOS ONE*, 14(12), e0226765. <https://doi.org/10.1371/journal.pone.0226765>
- Stark, G., Hart, G., Nartowt, B., & Deng, J. (2019). Predicting breast cancer risk using personal health data and machine learning models. *PLOS ONE*, 14(12), e0226765. <https://doi.org/10.1371/journal.pone.0226765>