INFORME CIENTIFICO DE BECA

Legajo N°:

TIPO DE BECA Doctoral 3año (ex BP16) **PERIODO** 04/2016 - 03/2017

1. DATOS PERSONALES

APELLIDO: Cravero

NOMBRES: Fiorella

Dirección Particular: Calle:

Localidad: Bahía Blanca CP: 8000 Tel:

Dirección electrónica (donde desea recibir información, que no sea "Hotmail"): fiorcravero7@gmail.com

2. TEMA DE INVESTIGACION (Debe adjuntarse copia del plan de actividades presentado con la solicitud de Beca)

"Modelado QSPR de Sistemas Complejos para Informática Molecular: desarrollo de Métodos de Selección y Extracción de Variables en presencia de incertidumbre"

PALABRAS CLAVE (HASTA 3)

+ Modelado de Materiales poliméricos polidispersos
+ QSPR
+ Selección de Características (Feature Selection)

3. OTROS DATOS (Completar lo que corresponda)

BECA DOCTORAL 1º AÑO (ex ESTUDIO 1° AÑO)**:** *Fecha inicio*:

BECA DOCTORAL 2º AÑO (ex ESTUDIO 2° AÑO)**:** *Fecha inicio*:

BECA DOCTORAL 3° AÑO (ex PERFECCIONAMIENTO 1º AÑO)**:** *Fecha inicio*: 04/2016

BECA DOCTORAL 4° AÑO (ex PERFECCIONAMIENTO 2º AÑO)**:** *Fecha inicio*:

4. INSTITUCION DONDE DESARROLLA LA TAREA

Universidad y/o Centro: Planta Piloto de Ingeniería Química (PLAPIQUI)

Facultad:

Departamento:

Cátedra:

Otros:

Dirección: Calle: La Carrindanga Nº: 7km

Localidad: Bahía Blanca CP: 8000 Tel: 0291-486-1700

5. CARGO UNIVERSITARIO *(si existe, especificar categoría, dedicación, condición de ordinario, regular o interino):*

6. CARGOS EN OTRAS INSTITUCIONES:

7. DIRECTOR DE BECA

Apellido y Nombres: Díaz, Mónica Fátima

Dirección Particular: Calle:

Localidad: Bahía Blanca CP: 8000 Tel:

Dirección electrónica: mdiaz@plapiqui.edu.ar

8. RESUMEN DE LA LABOR QUE DESARROLLA

Descripción para el repositorio institucional. Máximo 150 palabras.

De acuerdo a los objetivos de la tesis, la labor realizada se divide en 2 áreas de aplicacion: 1-Diseño de Nuevos Materiales poliméricos, 2-Diseño Racional de Drogas.
1.a-Representación Molecular Sintética (monomeros,trimeros): se avanzó en la prediccion de propiedades mecánicas con QSPR mediante Selección de características unido al Aprendizaje de caracteristicas. Se comenzó con modelos de Clasificación para ductilidad en polímeros.
1.b-Representación Molecular Real: no hay antecedentes de abordaje macromolecular para materiales poliméricos polidispersos. Se propuso una nueva metodología de modelado predictivo que considera: la recuperacion teórica de la Curva de distribución de pesos, Representantes Moleculares de diferentes tamaños, el desarrollo de un Graficador de macromoleculas y de un Calculador de descriptores (softwares inexistentes hasta el momento).
2.a-Idem 1.a para propiedades ADMET. En e-health, se modelaron por clasificación VOCs importantes en calidad de aire.
2.b- Se investigó el Dominio de Aplicación de los modelos QSPR, utilizando aprendizaje profundo y clasterización difusa.

9. EXPOSICION SINTETICA DE LA LABOR DESARROLLADA EN EL PERIODO.

Debe exponerse la orientación impuesta a los trabajos, técnicas empleadas, métodos, etc., y dificultades encontradas en el desarrollo de los mismos, en el plano científico y material. Si corresponde, explicite la importancia de sus trabajos con relación a los intereses de la Provincia.

El plan de trabajo presentado, persigue el desarrololo de una tesis doctoral. En el presente informe del periodo (04/2016-01/2017) se lograron realizar todas las tareas propuestas y se enumeran a continuación bajo títulos temáticos:

Avances en el desarrollo de metodologías predictivas

I- Hibridación de Feature Selection y Feature Learning

En Quimioinformática, uno de los campos más desarrollados es el modelado QSAR/QSPR (Relación Cuantitativa Estructura-Actividad/Propiedad). Para la construcción de un modelo QSAR/QSPR, se requiere una tabla de datos (descriptores moleculares), para luego elegir aquellas variables que mejor se ajusten, es decir que se aplica un método de Selección de Características o Aprendizaje de Características (también conocido como Extracción de Características). En todos los métodos de Selección de Características (Feature Selection), las variables se introducen en el modelo de manera algorítmica y una función de aptitud o criterios de selección determina qué variable debe retenerse o eliminarse del modelo. En el caso de los métodos de Aprendizaje de Características (Feature Learning), el conjunto original de descriptores se proyecta en nuevas variables en un espacio dimensional reducido, sin pérdidas de información. De la revisión bibliográfica, hemos podido concluir que siempre se presentan como técnicas competitivas, es decir por separado. Entonces, propusimos una estrategia que permite la confección de modelos QSPR de predicción combinando ambas técnicas, métodos de Selección de Características y Aprendizaje de Características, esperando una mejora en la performance. Se aplicó sobre dos campos: I-A-polímeros (visión sintética) y I-B- Drogas.

I-A: Esta nueva idea se utilizó para la predicción de la Resistencia a la Rotura para polímeros de alto peso molecular [ver publicaciones en 10.1.A.1], obteniendo buenos e interesantes resultados, a los que luego se los analizó mediante técnicas de analítica visual para estudiar y comprender mejor los resultados obtenidos [ver publicaciones en 10.1.A.2]. Mediante esta estrategia híbrida se obtuvieron mejores resultados que con Aprendizajes de características y resultados similares que con Selección de Características.

I-B: Por otro lado, se aplicó la combinación de estas técnicas a la predicción de actividad biológica para Barrera Sangre-Cerebro (BBB), absorción intestinal humana (HIA) y Exceso Enantiomérico (EE) en el diseño racional de fármacos, estrategia que fue desarrollada en un contexto de analítica visual [ver publicaciones en 10.3.A.1].

Conclusión: En ambos campos (polímeros y fármacos) obtuvimos modelos con buen rendimiento estadístico; aunque en polímeros es menor. Esto se condice con los resultados que se obtuvieron en una primera instancia con la técnica de Aprendizaje de Características para polímeros, y que puede deberse a la variabilidad estructural que presenta la base de datos con la que se trabaja. De aquí, la necesidad de mejorar la base de datos en términos de número y balance estructural.

II-Modelado Predictivo de materiales poliméricos polidispersos
En casos de estudio particulares, como es el diseño de nuevos materiales poliméricos, las moléculas involucradas son altamente complejas, caracterizadas por una distribución de pesos moleculares que deriva en una distribución probabilística de valores que puede adoptar un descriptor y no en un único valor para ese descriptor. Para abordar el ítem del modelado en presencia de variables con incertidumbre, se propone una estrategia de trabajo [ver publicaciones en 10.1.C.3] para realizar un modelado macromolecular de polímeros teniendo en cuenta la distribución de pesos moleculares de un material polimérico que consta de los siguientes ítems: II-A- Curvas, II-B- Representación de macromoléculas, II-C- QSPR para polidispersión

II-A: La polidispersión, es decir más de un único peso, es la característica que distingue a los polímeros sintéticos de otros materiales. Una de las principales dificultades que tiene esta estrategia es contar con la curva de polidispersión de pesos, ya que no es habitualmente reportada en la bibliografía. Por esta razón, fue necesaria la confección de un algoritmo de aprendizaje supervisado para poder predecirla y utilizarla. Para esto propusimos un método capaz de estimar la curva a partir de dos parámetros: Mn (Peso Molecular Promedio en número) y Mw (Peso Molecular Promedio en peso) del polímero del que quiere reconstruirse dicha curva. Fue necesario entrenar el algoritmo con curvas de distribución de pesos reales [ver publicaciones en 10.1.C.1].
Destaco que el principal inconveniente es contar con una base de datos de curvas reales lo suficientemente grande, variada y balanceada para que el algoritmo no sobreajuste y sus resultados sean confiables. Actualmente seguimos trabajando en mejorar dicha base de datos, esto es particularmente complicado ya que como se mencionó anteriormente las curvas no son publicadas en la literatura, y en la industria la mayoría de los polímeros de diseño están bajo patentes por lo que no es posible acceder a su información.

II-B Representación de macromoléculas

II-B-i. Representantes: La estrategia que proponemos para el abordaje macromolecular es una vez reconstruida la curva, dividirla en 10 partes iguales y de cada una de estas fracciones obtener un representante (ej. peso molecular promedio de cada fracción). Los pesos moleculares de los representantes son muy altos y por lo tanto el número de Unidades Repetitivas (UR) necesarias para alcanzarlo es elevado, lo cual imposibilita el uso de softwares comerciales utilizados hasta el momento (abordaje sintético) como HyperChem

para graficar y estabilizar dichas molécula y Dragón para calcular descriptores. Estas limitaciones impuestas por el uso de herramientas disponibles fueron reportadas [ver publicaciones en 10.1.C.2].
Es válido aclarar en este punto que dividir la curva en 10 representantes es sólo la primera aproximación. Se pretende, dividirla en más y menos fracciones, en busca de un óptimo. Así como trabajar con números promedios como la media o mediana.

II-B-ii. Polimerizador (representación de macromoléculas): Como una solución a la limitación de estos softwares (debido al altísimo peso molecular), se propuso un método de polimerización in silico [ver publicaciones en 10.1.B.1]. Este "polimerizador" se basa un algoritmo de representación mediante código SMILES. La conversión a SMILES se logra a partir de un archivo MOL del monómero. La característica fundamental de esta transformación es conservar la especificación de la cabeza y cola de la UR y a partir de esto, es posible el desarrollo de un algoritmo que imita una polimerización cabeza-cola.
Resumiendo, la principal ventaja de este método es la posibilidad de obtener moléculas tan grandes como se desee ya que al trabajar con cadenas de caracteres (código SMILES) no es demandante a nivel recursos computacionales, el límite lo impone el espacio en memoria que se disponga. Aún no hemos establecido el límite en la busca de los pesos moleculares que necesitamos conseguir (1x107 g/mol y más). Otra ventaja es que al trabajar con SMILES, es decir 2D, no es necesaria la estabilización de las moléculas lo cual deriva en un ahorro de tiempo importante y también de recursos computacionales.

II-B-iii. Cálculo de descriptores: El desafío es poder calcular descriptores moleculares a estas nuevas (y enormes) moléculas generadas. La principal dificultad es que, mediante el uso de paquetes y librerías específicas para el cálculo de descriptores, a medida que las moléculas crecen (en cantidad de UR) el número de descriptores posibles de calcular desciende [ver publicaciones en 10.3.B.1]. Solucionar este problema es un trabajo en progreso. Entonces, identificar del tipo de descriptores que no pueden calcularse es el primer paso, también se buscan nuevas librerías de cálculo de descriptores en general y en particular del tipo que no pueden resolver las anteriores, para así conseguir un número adecuado de descriptores moleculares.
Destaco que la principal dificultad con la que nos hemos encontrado en este punto es que los paquetes de cálculo tienen un límite que aún no hemos podido identificar con precisión. Este límite está relacionado con el peso molecular, por lo que se buscan nuevas librerías de cálculo de descriptores open source, es decir de código abierto para poder tratar de salvar este límite.

II-C- QSPR para polidispersión
Una vez que se cuenta con los descriptores para múltiples instancias de peso molecular de un polímero (representantes), es necesario abordarlos conjuntamente. Para esto se diseñó una estrategia basada en teoría de lógica difusa que se encuentra en etapa de diseño, que consta de la creación de una base de datos de descriptores moleculares representados por números difusos. El primer paso es la construcción de una curva acumulada de valores que toma cada descriptor, de modo de tener una curva por descriptor. Luego, esta curva es transformada a un número difuso. El ultimo paso es diseñar un método de Selección de Características que trabaje con estos números difusos, es decir que los pueda tener como dato de entrada.
Destaco que este es un trabajo en progreso y es totalmente novedoso, ya que no existe antecedentes de abordaje de este estilo para estas problemáticas en la literatura. Esta idea-proyecto se desarrolla en colaboración con el grupo de investigación del Dr. Carlos Barrancos, especialista en lógica difusa que trabaja en Intelligent Data Analysis (DATAi), división de Ciencias de la Computación de la Universidad Pablo de Olavide en Sevilla, España.

Conclusión: En la actualidad estamos trabajando en la representación más realista de polímeros mediante sus pesos promedios Mn y Mw, enfoque macromolecular y en el

cálculo de varios tipos de descriptores clásicos, con el objetivo de conseguir un número elevado de descriptores que asegure que brinden información global de la estructura química de las moléculas, para luego poder proponer modelos de predicción QSPR de propiedades mecánicas de dichos polímeros que superen ampliamente al enfoque micromolecular. El desafo reside en el desarrollo de métodos Selección de Características para polidispersión.

III- Modelos de Clasificación.
Además de los Modelos de Regresión antes descriptos, se está trabajando en Modelos de Clasificación. Para la construcción de estos modelos se tiene en cuenta tanto descriptores clásicos, calculados por software comerciales como Hyperchem y Dragón, así como otros descriptores provenientes de enfoques de aprendizaje de características como CODES-Tsar. Se aplicó para fármacos en particular para Barrera Sangre-Cerebro (BBB), absorción intestinal humana (HIA) y Exceso Enantiomérico (EE), [ver publicaciones en 10.3.A.1].
También fue utilizada en el campo del diseño de materiales poliméricos para la caracterización de la ductilidad, clasificándolos en dúctil, frágil e indefinido [ver publicaciones en 10.3.B.2].
En el área de e-health, precisamente en lo que se refiere Compuestos Orgánicos Volátiles (VOCs) es interesante poder clasificarlos según su afinidad por tejido adiposo, por sangre o por igual afinidad entre ambos. En particular se trabajó con una base de datos de VOCs y a partir de su valor para LogPliver se los clasificó en esas tres clases diferentes [ver publicaciones en 10.2.A.1].

Conclusión: Los modelos de clasificación responden a una necesidad de mayor comprensión de la salida (resultado) de los modelos, ya que algunas veces es más importante clasificar a las moléculas y no conocer en detalle el valor con el que se relaciona cierta molécula con la propiedad target. Como estas son primeras aproximaciones, estamos en proceso de mejora de los modelos así como la forma de evaluar los mismos.

IV- Dominio de Aplicación
Se conoce como dominio de aplicación al espacio fisicoquímico en el cual un modelo ha sido desarrollado y para el cual es posible hacer predicciones realistas de nuevos compuestos. En nuestro grupo se desarrolló un modelo que clasifica al nuevo compuesto en una de tres clases bien diferenciadas: confiable (S1), no confiable (S2) y no determinado (S3). Persiguiendo el objetivo de minimizar las moléculas clasificadas en S3, se ensayaron técnicas de lógica difusa (Fuzzy clasification) [ver publicaciones en 10.1.B.1] y de aprendizaje profundo (Deep Learning) [ver publicaciones en 10.1.B.2].

Conclusión: Como se puntualizó anteriormente, contar con una base de datos grande y balanceada es difícil de conseguir, y por lo tanto definir el dominio de aplicación, en la que una base de datos tiene incumbencia es de vital importancia.

10. TRABAJOS DE INVESTIGACION REALIZADOS O PUBLICADOS EN ESTE PERIODO.

***10.1 PUBLICACIONES**. Debe hacer referencia exclusivamente a aquellas publicaciones en la cual se haya hecho explícita mención de su calidad de Becario de la CIC (Ver instructivo para la publicación de trabajos, comunicaciones, tesis, etc.). Toda publicación donde no figure dicha mención no debe ser adjuntada ya que no será tomada en consideración. A cada trabajo asignarle un número e indicar el nombre de los autores, en el mismo orden en que aparecen en la publicación, informe o memoria técnica, lugar donde fue publicado, volumen, página y año si corresponde. En cada trabajo que el becario presente -si lo considerase de importancia- agregará una nota justificando el mismo y su grado de participación. Asimismo, en cada caso deberá indicar si el trabajo se encuentra depositado en el repositorio institucional CIC-Digital.*

A) Artículos en Revistas Internacionales con Referato Indexadas en Scopus (Sin Factor de Impacto en ISI):

1 - Cravero Fiorella, Martínez M. Jimena, Vázquez Gustavo E., Díaz Mónica F., Ponzoni Ignacio. “Intelligent Systems for Predictive Modelling in Cheminformatics: QSPR Models for Material Design using Machine Learning and Visual Analytics Tools”, Advances in Intelligent Systems and Computing, Vol. 477, pp. 3-11,(2016). Springer-Verlag. ISSN 2194-5357. (Artículo completo en conferencia). doi 10.1007/978-3-319-40126-3_1.

2- Cravero Fiorella, Martínez M. Jimena, Vázquez Gustavo E., Díaz Mónica F., Ponzoni Ignacio. “Feature Learning applied to the Estimation of Tensile Strength at Break in Polymeric Material Design”, Journal of Integrative Bioinformatics, 13 (2), 286 (2016). ISSN: 1613-4516. doi:10.2390/biecoll-jib-2016-286.

B) Publicaciones en Actas de Congresos Internacionales con referato

1 - Cravero Fiorella, Vázquez Gustavo E., Ponzoni Ignacio, Díaz Mónica F. “Modelado molecular de materiales poliméricos en quimioinformatica”. Presentación Oral. SAM-CONAMET 2016. Córdoba, Argentina. (Noviembre 2016).

2 - Cravero Fiorella, Martínez M. Jimena, Díaz Mónica F., Ponzoni Ignacio. “Fuzzy Clustering: Identification of Similar Compounds for Virtual Screening in Rational Drug Design”. The fourth International Society for Computational Biology Latin America Bioinformatics Conference (4º ISCB-LA). Buenos Aires, Argentina. (Noviembre 2016).

3 - Martínez M. Jimena, Cravero Fiorella, Díaz Mónica F., Ponzoni Ignacio. “Unsupervised Learning Based on Deep Learning Applied to the Identification of Applicability Domain of QSAR Models”. The fourth International Society for Computational Biology Latin America Bioinformatics Conference (4º ISCB-LA). Buenos Aires, Argentina. (Noviembre 2016).

C) Publicaciones en Actas de Congresos Nacionales:

1 - Cravero Fiorella, Martínez M. Jimena, Vázquez Gustavo E., Ponzoni Ignacio, Díaz Mónica F. “Prediccion de curvas teóricas de distribución de peso molecular de resinas polimericas” (5pag). 31° Congreso Argentino de Química, Buenos Aires, Argentina (Octubre, 2016).

2 - Cravero Fiorella, Martínez M. Jimena, Vázquez Gustavo E., Ponzoni Ignacio, Díaz Mónica F. “Representacion de la estructura molecular de polimeros sinteticos de alto peso” (5pag). 31° Congreso Argentino de Química, Buenos Aires, Argentina (Octubre, 2016).

3 - Cravero Fiorella, Ponzoni Ignacio, Díaz Mónica F. "Informática Molecular: Polímeros modelados por Distribución de Pesos" III Congreso Internacional de Ciencia y Tecnología de la Provincia de Buenos Aires" (CICyT CIC 2016), La Plata - Argentina, Septiembre 2016.

10.2 TRABAJOS EN PRENSA Y/O ACEPTADOS PARA SU PUBLICACIÓN*. Debe hacer referencia exclusivamente a aquellos trabajos en los que haya hecho explícita mención de su calidad de Becario de la CIC (Ver instructivo para la publicación de trabajos, comunicaciones, tesis, etc.). Todo trabajo donde no figure dicha mención no debe ser adjuntado porque no será tomado en consideración. A cada trabajo, asignarle un*

número e indicar el nombre de los autores en el mismo orden en que aparecen en la publicación y el lugar donde será publicado. A continuación, transcribir el resumen (abstract) tal como aparecerá en la publicación. La versión completa de cada trabajo se presentará en papel, por separado, juntamente con la constancia de aceptación. En cada trabajo, el becario deberá aclarar el tipo o grado de participación que le cupo en el desarrollo del mismo y, para aquellos en los que considere que ha hecho una contribución de importancia, deberá escribir una breve justificación.

A) Publicaciones en Actas de Congresos Internacionales con referato (trabajo extendido, con ISBN):

1 - Cravero Fiorella , Martínez M. Jimena, Díaz Mónica F., Ponzoni Ignacio. "Classification models for affinity to blood or liver prediction for volatile organic compounds using cheminformatics approaches" (8 pag). 5th International Work-Conference on Bioinformatics and Biomedical Engineering a realizarse en Granada, España (Abril 2017).

Abstract: In this work, we present Quantitative Structure-Activity Relationship (QSAR) classification models for characterization of molecules affinity to blood or liver for volatile organic compounds (VOCs), using information provided from log Pliver measures for VOCs. The models are computed from a dataset of 122 molecules. As a first phase, alternative subsets of relevant molecular descriptors related to the target property are selected by using feature selection methods and visual analytics techniques. From these subsets, several QSAR models are inferred by different machine learning methods. These models allow classifying a new compound as a molecule with affinity to blood, to the liver or equal affinity to both. The model with the highest performance correctly classifies 72.13% of VOCs and has an average receiver operating characteristic area equal to 0.83. As a conclusion, this QSAR model can predict the medium affinity of a VOC, which can help in the development of physiologically based pharmacokinetic computational models required in e-health.

***10.3 TRABAJOS ENVIADOS Y AUN NO ACEPTADOS PARA SU PUBLICACION**. Incluir un resumen de no más de 200 palabras de cada trabajo, indicando el lugar al que ha sido enviado. Adjuntar copia de los manuscritos.*

A) Artículos enviados a Revistas Internacionales con Referato Indexadas en ISI y en Scopus

1 - Ignacio Ponzoni, Victor Sebastian, Carlos Requena, Carlos Roca, María J. Martínez, Fiorella Cravero, Mónica F. Díaz, Juan A. Páez, Ramon Gomez-Arrayas, Javier Adrio and Nuria E. Campillo. "Hybridizing Feature Selection and Feature Learning Approaches in QSAR Modeling for Drug Discovery". Scientific Report (Nature), DICIEMBRE, 2016. ISSN: 2045-2322. NOTA: debido a requerimientos de copyright no se puede presentar el manuscrito completo hasta su aceptación.

QSPR modeling using machine learning techniques constitutes a complex computational problem, where the identification of the most informative molecular descriptors for predicting a specific target property plays a critical role. Two main general approaches can be used for this modeling procedure: feature selection and feature learning. In this paper, a performance comparative study of two state-of-art methods related to these two approaches is carried out. In particular, regression and classification models for three different issues are inferred using both methods under different experimental scenarios: two drug-like properties, such as blood-brain-barrier and human intestinal absorption, and enantiomeric excess, as a measurement of purity used for chiral substances. Beyond the contrastive analysis of feature selection and feature learning methods as competitive approaches, the hybridization of these strategies is also evaluated based on previous results obtained in material sciences. From the experimental results, it can be concluded that there is not a clear winner

between both approaches because the performance depends on the characteristics of the compound databases used for modeling. Nevertheless, it was observed that the accuracy of the models can be improved by combining both approaches when the molecular descriptor sets provided by feature selection and feature learning contain complementary information.

B) Publicaciones en Actas de Congresos Internacionales con referato (trabajos sin ISBN):

1 - Cravero Fiorella, Schustik Santiago, Martínez M. Jimena, Ponzoni Ignacio, Díaz Mónica F. “Macro Approach to Molecular Modelling of Linear Polymers Applied to Estimation of Tensile Modulus for New Materials Development”. MATERIAIS 2017 (VIII International Symposium on Materials), a realizarse en Aveiro, Portugal (Abril 2017).

There has been much progress in the knowledge of relationships between the molecular structure of a material and its properties that lead to the ability to predict material properties previous synthesis. However, it is not easy to achieve these predictions for polymers since the involved variables are very complex because of high molecular weight and polydispersity. To the best of our knowledge, this is one of the first attempts to investigate the possibility of predicting tensile properties for polymers by using QSPR techniques and considering real average MW, instead of monomer. At present work, main problems for modeling real polymeric materials are described and innovative solutions are proposed. Representing (2D) high MW polymeric molecules cannot be done with available software. We developed an algorithm based on SMILES code that reaches 1x107 g/mol and more. On the other hand, is requires calculating molecular descriptors which is also not possible with available software. In this stage, some descriptors were calculated with different packages of the R programming language. Once these problems were solved, we applied methodology used for micro approach: A-Feature Selection, B-Computational Model and C-Physic-chemical interpretation. The obtained results are promising and the models look more robust than micro approach ones.

2 - Martínez M. Jimena, Cravero Fiorella, Díaz Mónica F., Ponzoni Ignacio. “QSPR Modeling Applied to High Molecular Weight Polymers: Ductility Characterization from Elongation at Break”. MATERIAIS 2017

New polymeric materials with specific requirements are designed to satisfy a demanding market. QSPR-models estimate a target property of chemical compounds from variables that describe their molecular structure. Here, we present classification QSPR-models for ductility characterization of polymers, using information provided by tensile test experiments in order to classify polymer ductility based on elongation at break. The models are computed from a dataset of 77 linear amorphous thermoplastic polymers with high molecular weight. The first step detects alternative subsets of relevant molecular descriptors related to the target property using a feature selection method. These subsets are contrasted by an expert via a visual analytics software tool. From this analysis, we conclude that an experimental parameter of the tensile test, cross-head-speed, plays a central role in all models. Finally, QSPR-models are inferred by different machine learning methods. The output of these models allows classifying a new virtual material under design as ductile, fragile or undefined. The model with the highest performance correctly-classifies 88.46% (%CC) of polymers and has a receiver-operating-characteristic (ROC) curve equal to 0.97. As conclusion, this QSPR model can predict if a material will be ductile or not in early phases of polymer design, previous synthesis, with high confidence.

10.4 TRABAJOS TERMINADOS Y AUN NO ENVIADOS PARA SU PUBLICACION*. Incluir un resumen de no más de 200 palabras de cada trabajo.*

10.5 COMUNICACIONES*. Incluir únicamente un listado y acompañar copia en papel de cada una. (No consignar los trabajos anotados en los subtítulos anteriores).*

1- "Highlights of the 1st Argentine Symposium of Young Bioinformatics Researchers (1SAJIB) organized by the ISCB RSG-Argentina" (8 Pag.) R. Gonzalo Parra, Lucas A. Defelipe, A. Brenda Guzovsky, Alexander M. Monzon, Fiorella Cravero, Estefanía Mancini, Nicolás Moreyra, Carla L. Padilla Franzotti, M. Victoria Revuelta, Maria I. Freiberger, Nahuel N. Gonzalez, German A. Gonzalez, Facundo Orts, Nicolas Stocchi, Marcia A. Hasenahuer, Elin Teppa, Diego J. Zea, Nicolas Palopoli. PeerJ Preprints (2016)

10.6 INFORMES Y MEMORIAS TECNICAS*. Incluir un listado y acompañar copia en papel de cada uno o referencia de la labor y del lugar de consulta cuando corresponda. Indicar en cada caso si se encuentra depositado en el repositorio institucional CIC-Digital.*

11. PUBLICACIONES Y DESARROLLOS EN:

11.1 DOCENCIA

11.2 DIVULGACIÓN

11.3 OTROS

En cada caso indicar si se encuentran depositados en el repositorio institucional CIC-Digital.

12. PARTICIPACION EN REUNIONES CIENTIFICAS. *Indicar la denominación, lugar y fecha de realización, tipo de participación que le cupo, títulos de los trabajos o comunicaciones presentadas y autores de los mismos.*

--"III Congreso Internacional de Ciencia y Tecnología de la Provincia de Buenos Aires" (CICyT CIC), La Plata - Argentina, Septiembre 2016.

Trabajo Presentado:

1- "Informática Molecular: Polímeros modelados por Distribución de Pesos"

--"XXXI Congreso Argentino de Química" (CAQA 2016), Ciudad Autónoma de Buenos Aires - Argentina, Octubre 2016.

Trabajos Presentado:

1- "Representación de la Estructura molecular de Polímeros Sintéticos de alto Peso"

2- "Predicción de Curvas Teóricas de Distribución de Peso Molecular de Resinas Poliméricas"

--"4th International Society for Computational Biology Latin America Bioinformatics Conference " (ISCB-LA 2016), Ciudad Autónoma de Buenos Aires - Argentina, Noviembre 2016.

Trabajos Presentado:

1- "Fuzzy Clustering: Identification of Similar Compounds for Virtual Screening in Rational Drug Design"

2- "Unsupervised Learning Based on Deep Learning Applied to the Identification of Applicability Domain of QSAR Models"

13. **CURSOS DE PERFECCIONAMIENTO, VIAJES DE ESTUDIO, ETC**. *Señalar características del curso o motivo del viaje, período, instituciones visitadas, etc, y si se realizó algún entrenamiento.*

Curso de Posgrado:

- "Machine Learning aplicado a Bioinformatica", Departamento de Ciencias e ingeniería de la Computacion. Profesores: Dres C. Barranco González - I. Ponzoni

- "Métodos de Investigación para Ciencias de la Computación", Departamento de Ciencias e ingeniería de la Computacion. Profesor: Dr. Carlos Ivan Chesnevar.

Seminario de Formación Superior de la PLAPIQUI

- "The Increasing Scope of Optimization in the Oil and Gas Industry". Agosto de 2016. Prof. Ignacio Grossman Universidad de Carnegio-Mellon, EEUU.

- "Ejemplos de Investigación en el Laboratorio de Química Agro-Industrial de Toulouse: una directiva eco-responsable". Agosto de 2016. Prof. Carlos Vaca García. Instituto Nacional Politécnico Toulouse, Francia.

- "Tamices moleculares aplicados a Catálisis: Avances y Desafíos". Diciembre de 2016. Prof. Sibele Berenice Castella Pergher. Universidad Federal de Río Grande Del Norte (UFRGN), Brasil.

Seminario auspiciado por la Dirección de Vinculación Tecnológica de CONICET

- "Public Speaking: Impacto en la comunicación científica y de divulgación". Diciembre de 2016. Profesor: Fernando Johann

Workshop:

- "Workshop on Data Visualization" Workshop organizado por el ISCB Student Council (ISCB-SC) y el Regional Student Group (RSG-Argentina). Noviembre de 2016 en la Universidad Nacional de San Martín. Profesor: Sean O'Donoghue. Garvan Institute of Medical Research in Sydney, Australia.

14. **SUBSIDIOS RECIBIDOS EN EL PERIODO**. *Indicar institución otorgante, fines de los mismos y montos recibidos.*

15. **DISTINCIONES O PREMIOS OBTENIDOS EN EL PERIODO**.

16. **TAREAS DOCENTES DESARROLLADAS EN EL PERIODO**. *Indicar el porcentaje aproximado de su tiempo que le han demandado.*

17. **OTROS ELEMENTOS DE JUICIO NO CONTEMPLADOS EN LOS TITULOS ANTERIORES**. *Bajo este punto se indicará todo lo que se considere de interés para la evaluación de la tarea cumplida en el período.*

- Miembro del Comité Organizador del 2do Simposio Argentino de Jóvenes Investigadores en Bioinformática (2SAJIB), llevado a cabo en mayo de 2016 en la Universidad de Buenos Aires (ver ítem 10.5. COMUNICACIONES)

- Miembro del Comité Organizador del 2nd Latin America Student Council Symposium (LASCS 2016), llevado a cabo en noviembre de 2016 en la Universidad Nacional de San Martín.

- Coodinación de Actividades Satélites, Workshops y Turoriales del fourth International Society for Computational Biology Latin America Bioinformatics Conference (ISCB-LA), llevado a cabo en noviembre de 2016 en la Universidad Nacional de San Martín.

- Presidente del Grupo de Estudiantes de Bioinformática y Biología Computacional de Argentina (RSG-Argentina). El RSG-Argentina (Regional Student Council) es miembro parte del Student Council (SC) perteneciente a la International Society for Computational Biology (ISCB). Período: 12/2016 - 06/2018

18. DESCRIPCION DEL AVANCE EN LA CARRERA DE DOCTORADO.
Debe indicarse los logros alcanzados en la carrera de Doctorado en relación a los requisitos particulares de la misma (cursos, seminarios, trabajos de campo, etc), así como el porcentaje estimado de avance en la tesis.

Cursos: COMPLETO

Avance de la Tesis: 60% aproximadamente

19. TITULO Y PLAN DE TRABAJO A REALIZAR EN EL PROXIMO PERIODO. *Deberán indicarse claramente las acciones a desarrollar.*
TEMA: “Modelado QSPR de sistemas complejos para informática molecular: desarrollo de métodos de selección y extracción de variables en presencia de incertidumbre”.

Se adjunta plan de trabajo, donde se detallan las tareas a realizar y su cronograma.

..... Firma del Director

……..... Firma del Becario

Condiciones de Presentación

A. El Informe Científico deberá presentarse dentro de una carpeta, con la documentación abrochada y en cuyo rótulo figure el Apellido y Nombre del Becario, la que deberá incluir:

a. Una copia en papel A-4 (puntos 1 al 14).
b. Las copias de publicaciones y toda otra documentación respaldatoria, deben agregarse al término del desarrollo del informe
c. Informe del Director de tareas con la opinión del desarrollo del becario (en sobre cerrado).

Nota: El Becario que desee ser considerado a los fines de una prórroga, deberá solicitarlo en el formulario correspondiente, en los períodos que se establezcan en los cronogramas anuales.