# Open Source Chemoinformatics Software including KNIME Analytics

# 54

Georgios Leonis, Georgia Melagraki, and Antreas Afantitis

## Contents

## Abstract

In this chapter, we present a brief description of compound datasets and programs developed to serve chemoinformatics as well as, more specifically, nanoinformatics purposes. Emphasis has been placed on publicly available tools and particularly on KNIME (Konstanz Information Miner), the most widely used freely available platform for data processing and analysis. Among a multitude

G. Leonis
Novamechanics Ltd, Nicosia, Cyprus

G. Melagraki (✉) • A. Afantitis (✉)
InSilicoLab LP, Athens, Greece
Novamechanics Ltd, Nicosia, Cyprus
e-mail: melagraki@novamechanics.com; melagraki@insilicolab.eu;
afantitis@novamechanics.com; afantitis@insilicolab.eu

J. Leszczynski et al. (eds.), *Handbook of Computational Chemistry*,
DOI 10.1007/978-3-319-27282-5_57

of studies that have demonstrated the usefulness of chemoinformatics tools to chemical and medicinal applications, herein we present indicative cases of five successful KNIME-based approaches. The first two studies include the risk assessment of nanoparticles (NPs) through the Enalos InSilicoNano platform, namely, (1) the prediction of the toxicity of iron oxide NPs and (2) the cellular uptake prediction of computationally designed NPs with the aid of reliable quantitative nanostructure–activity relationships (QNAR) models. The third case study deals with the recognition of organic substances as corrosion inhibitors though the construction of predictive quantitative structure–property relationships (QSPR) models with Enalos KNIME nodes. Finally, two more cases are briefly described and involve the accurate prediction of yellow fever inhibitors from the ChEMBL database and the de novo design of compounds with the reaction vectors methodology. The aim of this work is to familiarize the interested reader with the freely available in silico tools in KNIME analytics platform and to demonstrate their value and effectiveness toward specific computational applications.

## Introduction

Chemoinformatics employs computational methods and information technology to deal with chemical problems (Leach and Gillet 2007). Current efforts in the field of drug discovery are particularly concerned with the handling of chemical structural information so that properties of a ligand are optimized to address the multiple demands of a potent drug (Brown 1998). Chemoinformatics mainly emerged due to the enormous amount of data that has been generated by recent drug discovery attempts, including high-throughput screening and combinatorial chemistry methodologies (Russo 2002). Chemoinformatics include several modeling approaches aiming at successful drug design. For instance, the development of quantitative relationships between the observed biological activities and the chemical structures through construction of quantitative structure–activity relationships (QSARs) models and the prediction of ligand-protein structures via docking approaches are among the most widely used techniques. It is important to highlight that the methods employed in chemoinformatics are usually developed to handle large sets of chemical structures and their different properties (usually referred as molecular descriptors) and thus should be appropriate for big data analysis.

Through the course of the years, the development of chemoinformatics approaches has been greatly assisted by data mining tools and open-source software. The recently released curated small-molecule databases are valuable assets for testing and validating chemoinformatics algorithms and tools (Hu and Bajorath 2012). There are many databases (either publicly available or proprietary) containing a number of chemical substances. The size of typical databases ranges between tens of thousands and millions of compound entries (Leach and Gillet 2007). The most well used public databases are PubChem (Wang et al. 2009), ZINC (Irwin et al. 2012), BindingDB (Liu et al. 2007), and

ChEMBL (Gaulton et al. 2012). ZINC contains the three-dimensional structures of commercially available compounds, which were constructed to be used in structure-based virtual screening, while PubChem, BindingDB, and ChEMBL also include (bio)activity information. Additionally, PubChem provides screening information data sets (SIDS) and the three-dimensional structures of the majority of the compounds, and in BindingDB and ChEMBL databases the activities of compounds can be assembled into relevant classes. Activity classes are of particular usefulness toward benchmarking of new computational methodologies (Hu and Bajorath 2012).

Recently, virtual databases have been built. Virtual compounds are substances that have not been observed/synthesized so far, but they could be synthesized. This broader consideration allows for the construction of even larger virtual libraries containing billions of compounds (Leach and Gillet 2007).

The release of these important sources of information that can be systematically explored have boosted the development of many software tools for chemoinformatics that make extensive use of datasets. Among others, software tools have been particularly emerged to tackle research practices, such as data mining, virtual screening and machine learning, molecular selectivity analysis, and visualization of structure-activity relationships (SARs).

In this chapter, we briefly review some of the most popular chemoinformatics tools with particular emphasis on programs that are either publicly available or at least free for academic purposes. Since KNIME is the most widely used, freely available platform for chemoinformatics applications, we will specifically present tools that have been integrated into KNIME and are offered as KNIME nodes to execute several important tasks for chemoinformatics analysis. Such applications include CDK, Indigo, RDKit, Vernalis, CACTVS, Enalos, Lhasa, OpenBabel, OCHEM, Chemical Identifier Resolver, ErlWood, EMBL-EBI Nodes, and CheS-Mapper.

## Popular Software Tools for Chemoinformatics

Konstanz Information Miner (KNIME) is an open-source analytics platform, which is the leading tool for wide-ranging data processing, integration, analysis, and exploration (Berthold et al. 2008). It enables the visual creation of data flows (so-called pipelines), the selective execution of specified analysis steps, and the presentation of the results via interactive views on models and data. KNIME offers intuitive use and high level of scalability, which currently render it the most popular platform for chemoinformatics applications. Therefore, as already mentioned, this chapter will be mostly devoted to the description of software tools that employ KNIME to accomplish their functions.

The Chemistry Development Kit (CDK) is an open-source and development chemoinformatics software (Steinbeck et al. 2003). In collaboration with the KNIME group, the CDK nodes for KNIME have been recently developed. These nodes provide features regarding chemical compound handling, such as several file conversion applications for molecules, calculation and drawing of 2D and 3D structures, symmetry group calculations, fingerprint calculation, proper handling of hydrogen atoms, and molecular property estimations, among others.

Indigo (developed by GGA Software Services LLC) is a software tool for organic chemistry (http://lifescience.opensource.epam.com/indigo/). Manipulation and functionality of organic structures with the Indigo nodes for KNIME can be obtained through conversions to Kekulé and aromatic states, handling of hydrogen, molecular properties generation, fingerprint comparison, R-group decomposition, stereochemistry calculation, and component separation. Additional functionalities include file conversions among SDF, SMILES, and CML formats; detection of drawing errors in structures; 2D structure generation; and structure matching.

RDKit (http://www.rdkit.org/) also provides chemoinformatics applications through KNIME, for instance, substructure filtering and searching, 2D and 3D structure generation, chemical reactions, molecular fingerprinting, salt separation from compounds, and R-group decomposition.

Vernalis Research (http://www.vernalis.com/research) employs its KNIME nodes to assist structure-based and fragment-based drug discovery. It includes several functionalities, such as flow control, PDB and sequence tools, I/O applications, matched molecular pairs, and fingerprint properties.

Lhasa Limited nodes for KNIME offer additional operations on the evaluation of binary classification models and table manipulation (http://www.lhasalimited.org/).

The OpenBabel chemoinformatics package (http://openbabel.org/) is primarily a file converter toolbox (KNIME chemistry nodes). It can also filter molecular files with SMARTS and has a wide applicability in analyzing molecular modeling and bioinformatics data (O'Boyle et al. 2011).

The Online Chemical Modeling Environment (OCHEM) is a platform, which aims to simplify the procedures for performing QSAR calculations (Sushko et al. 2011). This is achieved through combination of experimental results taken from a database and a modeling procedure. The database contains thousands of entries and a user-friendly application environment. The use of current KNIME nodes for OCHEM is restricted to some running predictions and the data export/import.

KNIME nodes development from ErlWood also offers interpretation and handling of structure-activity relationship data, as well as various compound viewing facilities (https://tech.knime.org/community/erlwood).

The European Bioinformatics Institute (www.ebi.ac.uk) provides the EMBL-EBI KNIME nodes, which make use of the Chemical Entities of Biological Interest (ChEBI) database to obtain database files via ChEBI IDs, substructure, or keyword searches. ChEBI is a publicly available depository of small chemical molecules.

The visualization node CheS-Mapper (Gutlein et al. 2012) employs the characteristics of small molecule structures to perform clustering according to feature similarity criteria. 3D structure depiction and embedding of compound datasets are also supported.

The Chemical Identifier Resolver (CIR, developed by the CADD group at the National Cancer Institute) nodes for KNIME enable the recognition of a chemical structure provided that an identifier is known (http://cactus.nci.nih.gov/chemical/

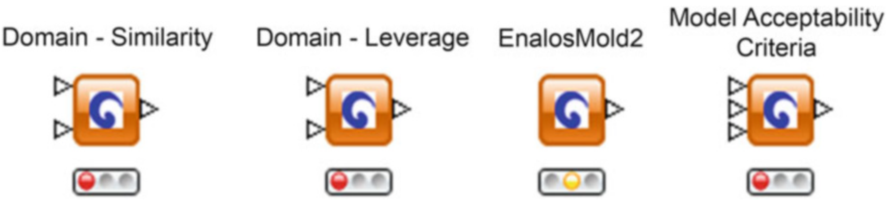


**Fig. 1** Freely available Enalos KNIME nodes

structure). CIR is a resolver for various structure identifiers and can also convert a particular structure identifier into another one.

Finally, the functionality of the Enalos KNIME nodes (https://tech.knime.org/community/enalos-nodes) developed by NovaMechanics Ltd will be briefly described. Enalos nodes for the KNIME platform are associated with several important aspects regarding data analysis and curation for chemoinformatics and nanoinformatics. The Enalos nodes (Fig. 1) among others provide (1) domain-similarity based on (i) euclidean distances or (ii) leverages, (2) fit quality and predictive power of a QSAR model with the Model Acceptability Criteria node, (3) the newly developed fast generation of all possible substitutions of a lead compound, and (4) calculation of important molecular descriptors with the Mold2 node. Mold2 is able to evaluate large and diverse sets of molecular descriptors from two-dimensional chemical structure information (Hong et al. 2008). Comparison of Mold2 descriptors with descriptors calculated from commercial software on several published datasets showed that Mold2 descriptors yield models with higher quality than other packages and also produce sufficient structural information.

In the following section, we present three case studies of in silico approaches developed by our group that involve the utilization of Enalos KNIME nodes for building predictive models.

## Chemoinformatics Studies Using the Enalos KNIME Nodes and Enalos Cloud Platform

The KNIME analytics platform contains a multitude of processing nodes for data I/O, modeling, analysis, and data mining. It integrates well with Weka programming language for machine learning applications. Over the last 30 years, a vast amount of data has been generated within the areas of chemo-, bio-, and nanoinformatics. KNIME has emerged as one of the most reliable open-source data mining tools for the prediction of chemical properties and applications, such as virtual screening of chemicals and nanoparticles (NPs), chemical library enumeration, virtual library creation, building QSAR/QSPR, ADMET, and pharmacokinetics models, as well as prediction of various biological effects of organic compounds and NPs.

In this section, we present three case studies that involve the use of Enalos KNIME nodes in predicting (i) the toxicity of iron oxide NPs, (ii) the cellular uptake of organic NPs (virtual screening is also demonstrated), and (iii) the inhibitory potency of organic substances against corrosion.

In the first two case studies Enalos InSilico platform is also introduced. Enalos InSilico platform is a Cloud platform built to host a variety of predictive models to address the need for risk assessment and virtual screening. Workflows included in Enalos InSilico platform are constructed based on reliable data information, and each workflow combines advanced in silico tools to yield accurate predictions. Predictive workflows are available in a user-friendly format and include toxicity, biological activity, and property evaluation models.

## Risk Assessment Tool for the Toxicity Prediction of Iron Oxide Nanoparticles Through Enalos InSilicoNano Cloud Platform

Nanoparticles (NPs) are known for their unique optical, electronic, and mechanical properties, which have led to the rapid evolution of nanotechnology materials being applied in a wide range of commercial, technological, and therapeutic applications in fields such as environment, industry, defense, electronics, and biomedicine. The latter has enjoyed great scientific, technological, and commercial progress in different NP applications (Gajewicz et al. 2012; Cohen et al. 2013).

Along with the apparent increasing use of NPs, concerns on their effect upon the environment and human health have been raised. Since toxicity assessment of NPs via traditional experimental routes often require expensive and time-consuming procedures, computational approaches such as quantitative nanostructure–activity relationships (QNARs) have been successfully used to predict the toxic effects of NPs (Vrontaki et al. 2015; Kleandrova et al. 2014a, b; Speck-Planche et al. 2015; Winkler et al. 2013, 2014; Shao et al. 2013; Toropov et al. 2013). However, the computational investigation of NP toxicity is seriously hindered by the lack of available NP descriptors, organized datasets, and systematic experimental data for NPs. Only few organized datasets on NP toxicity are available so far. Among these toxicity datasets on nanostructures, metal oxide NP data have been investigated in several computational studies (Fourches et al. 2010; Liu et al. 2011, 2013a; Puzyn et al. 2011; Zhang et al. 2012).

A fully validated QNAR model is presented, which was constructed based on toxicity data of iron oxide NPs with different core, coating, and surface modifications (Melagraki and Afantitis 2015; Shaw et al. 2008; Liu et al. 2013). The initial dataset was constructed with 44 iron oxide NPs that comprised a core with either $Fe_2O_3$ or $Fe_3O_4$ coating, including cross-linked dextran, PVA, or other, and various surface modifications (Shaw et al. 2008; Liu et al. 2013). Values for descriptors such as the size, R1 and R2 relaxivities, and zeta potential along with a coating-indicative parameter were considered as independent variables for the model development. The values of the input variables for each NP and the corresponding toxicity class are shown in Table 1.

**Table 1** NP properties, bioactivity, and predictions

| ID | NP[b] | Size | ZP | R1[c] | R2 | Coating | NHit[d] | Class | Prediction | Domain |
|---|---|---|---|---|---|---|---|---|---|---|
| 1[a] | NP1 | 36 | −19.9 | 19 | 45 | Cross-linked dextran | 1 | Inactive | Inactive | Reliable |
| 2[a] | NP2 | 30 | −9.22 | 26 | 74 | Cross-linked dextran | 1 | Inactive | Inactive | Reliable |
| 3 | NP3 | 32 | 5.9 | 21 | 54 | Cross-linked dextran | 3 | Inactive | Inactive | – |
| 4 | NP4 | 74 | −2.72 | 21 | 153 | Cross-linked dextran | 2 | Inactive | Inactive | – |
| 5 | NP5 | 27 | 3.34 | 17 | 36 | Cross-linked dextran | 0 | Inactive | Inactive | – |
| 6[a] | NP6 | 29 | 1.95 | 22 | 51 | Cross-linked dextran | 2 | Inactive | Inactive | Reliable |
| 7[a] | NP7 | 38 | −10.1 | 21 | 62 | Cross-linked dextran | 1 | Inactive | Inactive | Reliable |
| 8 | NP8 | 33 | −19.5 | 22 | 49 | Cross-linked dextran | 0 | Inactive | Inactive | – |
| 9 | NP9 | 36 | −14 | 19 | 45 | Cross-linked dextran | 3 | Inactive | Inactive | – |
| 10 | NP10 | 28 | 3.24 | 19 | 39 | Cross-linked dextran | 1 | Inactive | Inactive | – |
| 11 | NP11 | 31 | −9.46 | 23 | 59 | Cross-linked dextran | 4 | Inactive | Inactive | – |
| 12 | NP12 | 31 | 3.64 | 19 | 49 | Cross-linked dextran | 17 | Active | Active | - |
| 13[a] | NP14 | 28 | 2.34 | 19 | 39 | Cross-linked dextran | 4 | Inactive | Inactive | Reliable |
| 14 | NP15 | 24 | −11.7 | 22 | 54 | Cross-linked dextran | 1 | Inactive | Inactive | – |
| 15 | NP16 | 37 | 0.766 | 21 | 52 | Cross-linked dextran | 2 | Inactive | Inactive | – |
| 16 | NP17 | 38 | −20.7 | 21 | 62 | Cross-linked dextran | 3 | Inactive | Inactive | – |
| 17 | NP18 | 38 | −9.08 | 21 | 62 | Cross-linked dextran | 0 | Inactive | Inactive | – |
| 18 | NP19 | 31 | −3.61 | 19 | 49 | Cross-linked dextran | 8 | Active | Active | – |
| 19[a] | NP20 | 38 | −9.34 | 21 | 62 | Cross-linked dextran | 7 | Active | Inactive | Reliable |
| 20 | NP21 | 28 | −9.23 | 15 | 40 | Cross-linked dextran | 4 | Inactive | Inactive | – |
| 21 | NP22 | 36 | −21.9 | 36 | 122 | Cross-linked dextran | 2 | Inactive | Inactive | – |
| 22 | NP23 | 31 | −6.11 | 20 | 45 | Cross-linked dextran | 3 | Inactive | Inactive | - |
| 23 | NP26 | 40 | −12 | 15 | 30 | PVA | 3 | Inactive | Active | - |
| 24[a] | NP27 | 40 | −3.77 | 15 | 30 | PVA | 0 | Inactive | Active | Reliable |
| 25 | NP28 | 40 | −7.57 | 15 | 30 | PVA | 5 | Active | Active | - |
| 26 | NP29 | 40 | 0.25 | 15 | 30 | PVA | 7 | Active | Active | - |
| 27 | NP30 | 40 | −6.05 | 15 | 30 | PVA | 5 | Active | Active | - |
| 28 | NP31 | 20 | −12.3 | 0.5 | 0.5 | PVA | 4 | Inactive | Active | - |
| 29 | NP32 | 20 | −4.22 | 0.5 | 0.5 | PVA | 8 | Active | Active | - |
| 30 | NP33 | 20 | −7.15 | 0.5 | 0.5 | PVA | 0 | Inactive | Active | - |
| 31[a] | NP34 | 20 | −4.3 | 0.5 | 0.5 | Other | 13 | Active | Active | Reliable |
| 32 | NP35 | 20 | −12.1 | 0.5 | 0.5 | PVA | 8 | Active | Active | - |
| 33 | NP36 | 20 | −15.6 | 0.5 | 0.5 | Other | 9 | Active | Active | - |
| 34 | NP37 | 20 | −16.1 | 0.5 | 0.5 | PVA | 5 | Active | Active | - |
| 35 | NP38 | 20 | −4.7 | 0.5 | 0.5 | PVA | 13 | Active | Active | - |
| 36 | NP39 | 20 | −6.47 | 0.5 | 0.5 | PVA | 9 | Active | Active | - |
| 37[a] | NP40 | 20 | −6.54 | 0.5 | 0.5 | PVA | 6 | Active | Active | Reliable |
| 38 | NP41 | 20 | −10.8 | 0.5 | 0.5 | Other | 2 | Inactive | Inactive | - |
| 39[a] | NP42 | 20 | −7.7 | 0.5 | 0.5 | PVA | 6 | Active | Active | Reliable |
| 40[a] | NP43 | 20 | −6.75 | 0.5 | 0.5 | PVA | 6 | Active | Active | Reliable |

*(continued)*

**Table 1** (continued)

| ID | NP[b] | Size | ZP | R1[c] | R2 | Coating | NHit[d] | Class | Prediction | Domain |
|---|---|---|---|---|---|---|---|---|---|---|
| 41 | NP45 | 23 | −13.6 | 29 | 62 | Other | 1 | Inactive | Inactive | - |
| 42 | NP46 | 33 | −14.5 | 36 | 106 | Other | 1 | Inactive | Inactive | - |
| 43[a] | NP47 | 28 | −9.23 | 32 | 60 | Other | 0 | Inactive | Inactive | Reliable |
| 44[a] | NP48 | 25 | −37 | 29 | 49 | Other | 1 | Inactive | Inactive | Reliable |

[a] Test Set

[b] The 44 NPs were obtained from the following studies (Liu et al. 2011; Epa et al. 2012)

[c] R1: spin–lattice relaxivity; R2: spin–spin relaxivity

[d] NHit: the number of hits identified for each NP across the 64 bioactivity measures (4 cell lines X 4 assays X 4 concentrations)

NPs were evaluated with different assays in various cell types and concentrations that produced a 64-component vectorial metric. Each NP was characterized either active or inactive according to the number of hits obtained across the 64 bioactivity measures. For the QNAR developed KNIME workflow, the following steps have been integrated: (1) data preprocessing, (2) variable selection, (3) model development and (4) validation, and (5) determination of domain of applicability [via the Enalos Domain – Similarity node that defines applicability domain (APD) based on euclidean distances]. The publicly available set of Enalos KNIME nodes can be accessed through either the KNIME Community or NovaMechanics website (www.novamechanics.com/knime.php or www.insilicotox.com/index.php/products/enalos-knime-nodes-community-contributions/) (Melagraki and Afantitis 2013).

Before model running, the available data were separated into training set and test set with the partitioning node in KNIME. According to the data in the training set, the most significant descriptors were selected (Witten et al. 2005; Hall et al. 2009). The NPs used in the validation set were not further employed during model development. Among all available techniques, the J48 modeling method yielded the most predictively powerful model for the available data.

The proposed predictive model was validated internally and externally regarding goodness-of-fit, robustness, and predictivity, thus totally meeting the criteria recommended by the Organization for Economic Cooperation and Development (OECD).

To validate the performance of the model, the following parameters were considered (Afantitis et al. 2011; Mouchlis et al. 2012):

$$\text{Precision} = \text{TP}/\left(\text{TP} + \text{FP}\right)$$

$$\text{Sensitivity} = \text{TP}/\left(\text{TP} + \text{FN}\right)$$

$$\text{Specificity} = \text{TN}/\left(\text{TN} + \text{FP}\right) \text{ and}$$

$$\text{Accuracy} = \left(\text{TP} + \text{TN}\right) / \left(\text{TP} + \text{FP} + \text{FN} + \text{TN}\right),$$

where TP true positive, FP false positive, TN true negative, and FN false negative.

The confusion matrix is presented below:

| | Positive predicted | Negative predicted |
|---|---|---|
| Positive observed (Active) | TP | FN |
| Negative observed (Inactive) | FP | TN |

**Table 2** Confusion matrix (training set)

| | Positive predicted | Negative predicted |
|---|---|---|
| Positive observed (Inactive) | 16 | 0 |
| Negative observed (Active) | 4 | 11 |

**Table 3** Confusion matrix (test set)

| | Positive predicted | Negative predicted |
|---|---|---|
| Positive observed (Inactive) | 7 | 1 |
| Negative observed (Active) | 1 | 4 |

Additionally, the Y-randomization test demonstrated the robustness and the statistical significance of the predictive model.

The ability to perform virtual screening of NPs that were not originally included in the dataset is particularly important, especially if there is an indication on its reliability. For this purpose, it is crucial to determine the limits of the model's domain of applicability. This will identify NPs that are excluded from the area of reliable predictions of the proposed model.

In this work, the Enalos Domain-Similarity node calculated the domain of applicability using euclidean distances to estimate the similarity between NPs belonging to the training and test sets. More details on the domain of applicability calculation can be found elsewhere (Melagraki and Afantitis 2013; Afantitis et al. 2008; Tropsha 2010).

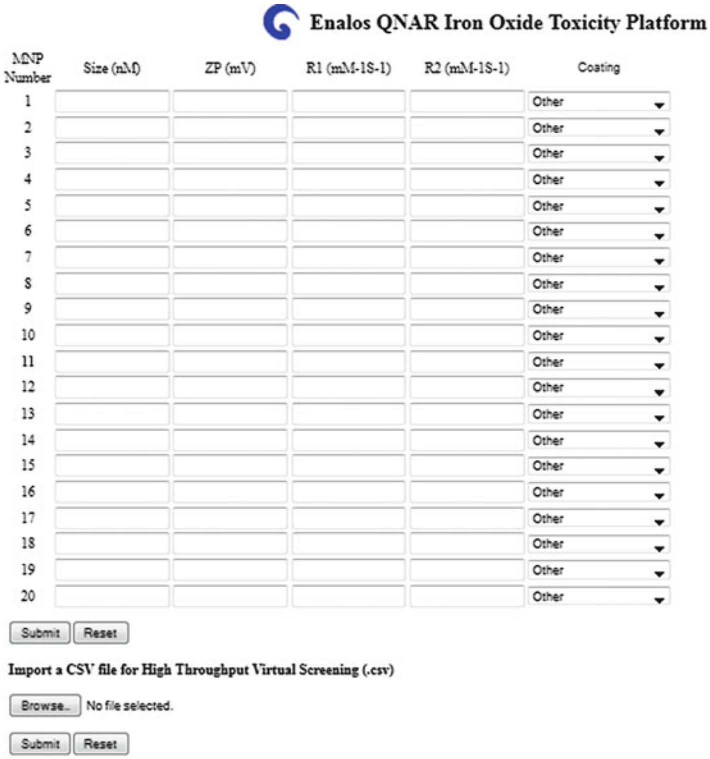
Based on the reported toxicity data for the 44 iron oxide NPs used in this study, 31 compounds were included in the training set while the other 13 in the test set. As mentioned above, subsequent model development was based on the NP structures of the training set. For the NPs, experimental parameters such as the size, R1 and R2 relaxivities, and zeta potential were combined with a coating-specific parameter and were used as inputs for the model development. The three coating categories which have been considered are: PVA, cross-linked dextran, and other. Among the above descriptors, the subset that best describes the variation of toxicity with NP properties (as demonstrated by the variable selection algorithm) includes the R1 and R2 relaxivities and coating.

After the model development from the training data was achieved, the toxicity prediction for the test set followed. Observed classes and predictions for the 44 NPs of the initial set are shown in Table 1. Confusion matrices for the training and the test sets are shown in Tables 2 and 3, respectively. The evaluation of the performance of the training set yielded: precision = 80 %, sensitivity = 100 %, specificity = 73.3 %, and accuracy = 87.1 %. Regarding the test set, the respective parameters are: precision = 87.5 %, sensitivity = 87.5 %, specificity = 80 %, and accuracy = 84.6 %.

The applicability domain was defined for NPs within the test set and the cutoff value was estimated to be 0.906. For all structures of the test set, values range from 0 to 0.448, therefore, all predictions are considered reliable. The above set of validation measurements highlights the accuracy, significance, and robustness of the produced model.

The predictive workflow is publicly available via the Enalos Cloud platform. In silico design and screening may be performed through the Enalos Cloud platform by visiting the iron oxide model web page (enalos.insilicotox.com/QNAR_IronOxide_Toxicity/). The user can initiate a prediction by either manually entering the selected NP properties (e.g., zeta potential, size etc.) or by importing a CSV file (.csv) with NP properties for high-throughput virtual screening (HTVS) (Scheme 1).

When properties are uploaded for a set of NPs and the input values have been included, the predictive model is used and a prediction is obtained. The generated



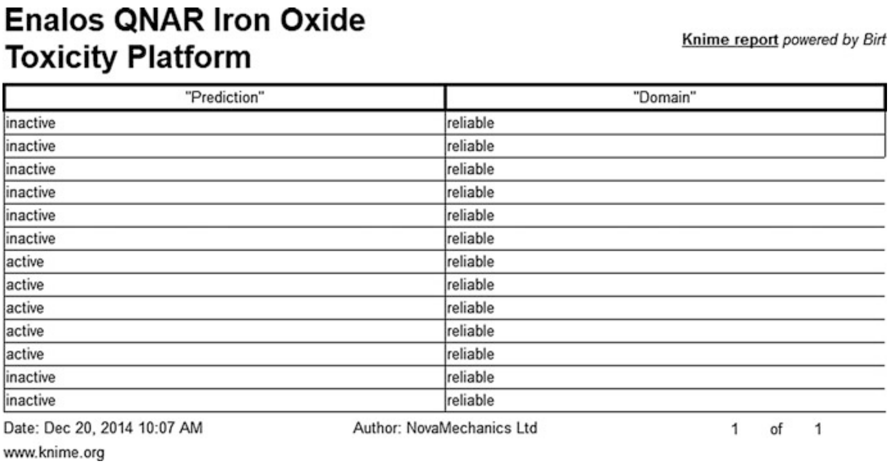
Enalos QNAR Iron Oxide Toxicity Platform

| MNP Number | Size (nM) | ZP (mV) | R1 (mM-1S-1) | R2 (mM-1S-1) | Coating |
|---|---|---|---|---|---|
| 1 | | | | | Other ▾ |
| 2 | | | | | Other ▾ |
| 3 | | | | | Other ▾ |
| 4 | | | | | Other ▾ |
| 5 | | | | | Other ▾ |
| 6 | | | | | Other ▾ |
| 7 | | | | | Other ▾ |
| 8 | | | | | Other ▾ |
| 9 | | | | | Other ▾ |
| 10 | | | | | Other ▾ |
| 11 | | | | | Other ▾ |
| 12 | | | | | Other ▾ |
| 13 | | | | | Other ▾ |
| 14 | | | | | Other ▾ |
| 15 | | | | | Other ▾ |
| 16 | | | | | Other ▾ |
| 17 | | | | | Other ▾ |
| 18 | | | | | Other ▾ |
| 19 | | | | | Other ▾ |
| 20 | | | | | Other ▾ |

Import a CSV file for High Throughput Virtual Screening (.csv)

No file selected.

**Scheme 1** Screenshot of Enalos iron oxide platform input page



Enalos QNAR Iron Oxide Toxicity Platform

Knime report *powered by Birt*

| "Prediction" | "Domain" |
|---|---|
| inactive | reliable |
| inactive | reliable |
| inactive | reliable |
| inactive | reliable |
| inactive | reliable |
| inactive | reliable |
| active | reliable |
| active | reliable |
| active | reliable |
| active | reliable |
| active | reliable |
| inactive | reliable |
| inactive | reliable |

Date: Dec 20, 2014 10:07 AM Author: NovaMechanics Ltd 1 of 1

www.knime.org

**Scheme 2** Screenshot of Enalos QNAR iron oxide toxicity platform results

output provides a summary of the results in a pdf-like format or a CSV file, which contains all relevant information for further analysis (Scheme 2). The results include predicted values for each nanostructure and a notification on the reliability of predictions based on the domain of applicability limits.

As already mentioned, the Enalos Cloud platform allows the performance of a preliminary in silico testing by virtually screening a set of NPs, based on the validated model. The predictive QNAR model described here can be accessed at the web page of Enalos QNAR Iron Oxide Toxicity Platform (http://enalos.insilicotox.com/QNAR_IronOxide_Toxicity/).

This study outlines Enalos InSilico platform as a useful tool to facilitate the computer-aided NP design by acting as a source of toxicity prediction for novel NPs.

## Cellular Uptake Prediction and Virtual Screening of Nanoparticles via Enalos InSilicoNano Platform

In this section, a validated quantitative nanostructure–activity relationships (QNAR) model that can predict the cellular uptake of organic nanoparticles is presented (Melagraki and Afantitis 2014). The model is publicly available through Enalos InSilico Cloud platform in the QNAR_PaCa2 web page (http://enalos.insilicotox.com/QNAR_PaCa2/) and can be used for new-structure predictions that are designed and/or uploaded to the server. The Enalos InSilico web service functionality was successfully used for the virtual screening of a set of PubChem structures, which were selected to recognize structures similar to that of a known active compound. The model was based on Mold2 descriptors and the k-nearest neighbors (kNN) algorithm.

The selected engineered NPs (ENPs) used for model development, possess the same metal core but different organic coating (Weissleder et al. 2005). The model building was based on a KNIME workflow, especially designed for this purpose. Initially, for the development of the model, all data including organic molecules and cellular uptake values were preprocessed and randomly divided into training set (89 compounds, which were used in the model development) and validation (25 compounds) set. The Enalos Mold2 KNIME node was used for the calculation of 777 descriptors for each compound. During the correlation analysis some descriptors were eliminated, thus leaving 382 of them to be used as inputs for the QNAR model development.

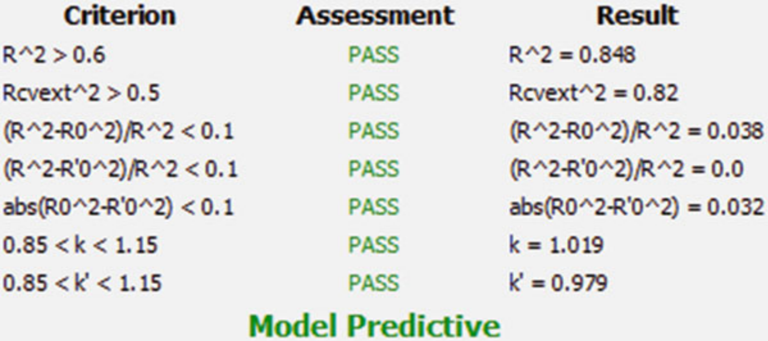
Next, the CfsSubset variable selection with BestFirst evaluator method was applied to identify nine descriptors as the most representatives of the structural features that define the biological profile of the studied NPs.

The nine descriptors are: (D461) Geary topological structure autocorrelation length-7 weighted by atomic van der Waals volume, (D467) Geary topological structure autocorrelation length-5 weighted by atomic Sanderson electronegativities, (D599) number of total quaternary C-sp3, (D649) number of group secondary aliphatic amines, (D712) number of group donor atoms for hydrogen bonds, (D714) number of group $CH_3R$ and $CH_4$, (D753) number of group phenol or enol or carboxyl OH, (D758) number of group $Al_2$–NH, and (D775) hydrophilic factor index. The physical meaning of the above descriptors can be found in the original publication (Melagraki and Afantitis 2014). The optimized value of k within kNN application was 2 (Franco-Lopez et al. 2001).

The proposed model was successfully validated with the methodology applied in the previous section. The validation results are shown in Fig. 2.

The $R^2_{LOO}$ was calculated to be 0.74. Also, the Y-randomization test verified the model's robustness and statistical significance. The decreased values of the correlation coefficient indicate the low possibility of chance correlation.

After the model was validated, the reliability of a given prediction was suggested through domain of applicability calculations (cutoff value = 2.153) (Mouchlis et al.



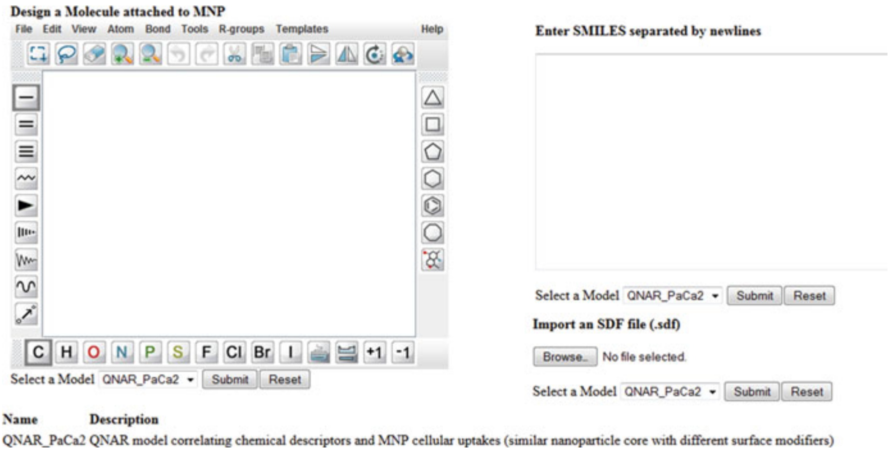
| Criterion | Assessment | Result |
|---|---|---|
| R^2 > 0.6 | PASS | R^2 = 0.848 |
| Rcvext^2 > 0.5 | PASS | Rcvext^2 = 0.82 |
| (R^2-R0^2)/R^2 < 0.1 | PASS | (R^2-R0^2)/R^2 = 0.038 |
| (R^2-R'0^2)/R^2 < 0.1 | PASS | (R^2-R'0^2)/R^2 = 0.0 |
| abs(R0^2-R'0^2) < 0.1 | PASS | abs(R0^2-R'0^2) = 0.032 |
| 0.85 < k < 1.15 | PASS | k = 1.019 |
| 0.85 < k' < 1.15 | PASS | k' = 0.979 |

Model Predictive

**Fig. 2** Model evaluation summary results

2012; Zhang et al. 1995; Papa et al. 2009). It was concluded that the proposed model requires only the structural information from the organic compounds involved and was confirmed to be accurate and reliable within applicability limits. Thus, the model could be considered a useful tool toward cellular uptake determination of NPs.

The model was made available online through Enalos InSilico platform for fast in silico predictions for a set of given compounds. Screenshots of the Enalos InSilico web service and the results page are presented in Schemes 3 and 4.



Design a Molecule attached to MNP

Select a Model QNAR_PaCa2

Enter SMILES separated by newlines

Select a Model QNAR_PaCa2

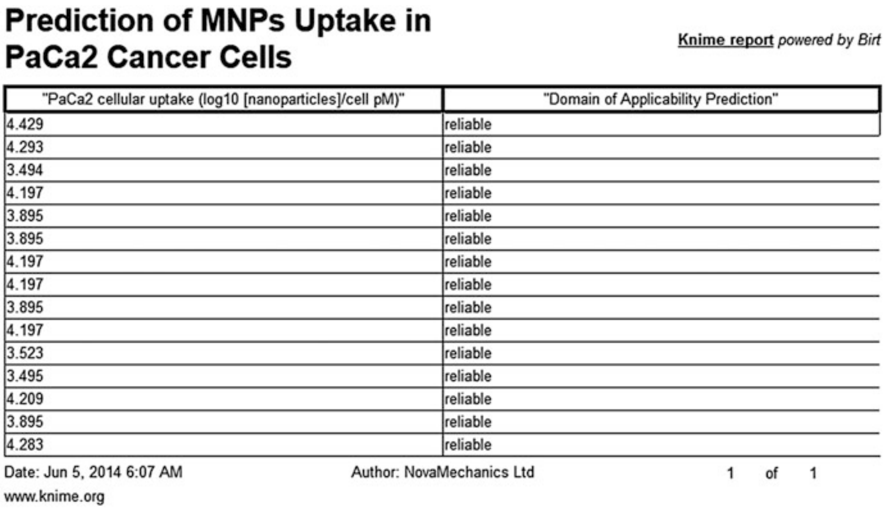
Import an SDF file (.sdf)

No file selected.

Select a Model QNAR_PaCa2

| Name | Description |
|---|---|
| QNAR_PaCa2 | QNAR model correlating chemical descriptors and MNP cellular uptakes (similar nanoparticle core with different surface modifiers) |

**Scheme 3** Screenshot of Enalos InSilico platform input page for QNAR_PaCa2 model



Prediction of MNPs Uptake in PaCa2 Cancer Cells

| "PaCa2 cellular uptake (log10 [nanoparticles]/cell pM)" | "Domain of Applicability Prediction" |
|---|---|
| 4.429 | reliable |
| 4.293 | reliable |
| 3.494 | reliable |
| 4.197 | reliable |
| 3.895 | reliable |
| 3.895 | reliable |
| 4.197 | reliable |
| 4.197 | reliable |
| 3.895 | reliable |
| 4.197 | reliable |
| 3.523 | reliable |
| 3.495 | reliable |
| 4.209 | reliable |
| 3.895 | reliable |
| 4.283 | reliable |

Date: Jun 5, 2014 6:07 AM Author: NovaMechanics Ltd 1 of 1

www.knime.org

**Scheme 4** Screenshot of Enalos InSilico platform results for QNAR_PaCa2 model

As shown in Schemes 3 and 4, the user can design or enter a chemical structure and obtain a prediction. The aforementioned workflow will calculate the descriptors and the output will be rapidly generated (within seconds). One may experiment with different scaffolds and structures and observe the structural features that induce a certain effect. Also, the user can exploit the proposed QNAR model and then scan specific structures for a preliminary in silico testing. In this way, it is possible to offer QSAR/QNAR results for immediate sharing and implementation. It was recently pointed out (Tetko 2012) that the use of predictive models as software tools will probably increase in the future, and this will activate the reuse of knowledge, which in turn will result in further developments.

As already mentioned, the proposed model and web platform can be used in virtual screening studies for the prioritization of new compounds. To demonstrate the usefulness of the produced model, potent compounds from the PubChem database were identified using similarity calculations based on molecular quantum numbers (MQNs) (Melagraki and Afantitis 2014). The virtual screening procedure was employed for the recognition of the first 1000 neighbors of **compound 36** (isatoic ahydride), in terms of chemical similarity. The 1000 resulting compounds were tested with the online tool through Enalos InSilico platform (via an sdf file, which contains all the structures) regarding their cellular uptake. Compounds were next classified by increasing potency and the most promising ones were selected for screening. The predictions for the first 20 compounds are shown in Table 4.

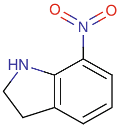
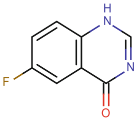
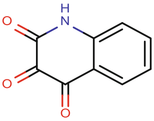
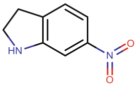
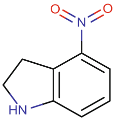
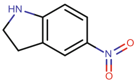
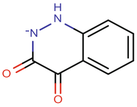
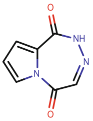
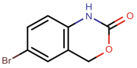
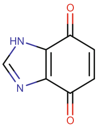
Within the proposed framework, the Enalos InSilico platform emerges as an invaluable application for the evaluation of novel NPs, which have not been experimentally tested or synthesized. An additional, important aspect of the approach is that the above tools can be further expanded and applied to polymer–NP structures that are currently gaining increasing attention.

## Identification of Organic Materials as Corrosion Inhibitors Based on Enalos KNIME Nodes

One of the most efficient ways to prevent metal corrosion in acidic media is the development of novel corrosion inhibitors (Ebenso et al. 2012). Organic inhibitors, which contain heteroatoms (e.g., oxygen, nitrogen, sulfur) and possess multiple bonds, have been considered for various corrosion systems, metals, and alloys. Inhibition is obtained with creation of physical and/or chemical absorption film on the surface of the metal (El Ashry et al. 2012). The planarity of heterocycles and the presence of lone electron pairs on the heteroatoms are crucial factors that control the absorption of these compounds on the metallic surface.

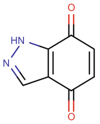
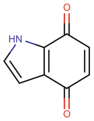
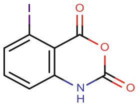
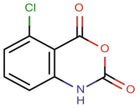
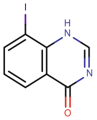
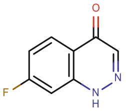
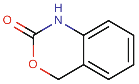
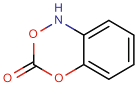
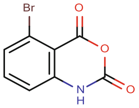
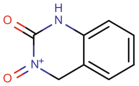
One drawback regarding the design of corrosion inhibitors is the time-consuming and costly nature of the process. On the other hand, computational techniques, such as the quantitative structure–property relationships (QSPR) methodology has greatly advanced the efficient modeling and prediction of new or modified corrosion inhibitors (Lee et al. 2012; Toropov et al. 2012).

**Table 4** Virtual screening results for the most promising compounds in PubChem database

| ID | Compound | Predicted value<br>PaCa2 cellular uptake<br>(log10 [NP]/cell) | Domain of applicability<br>(limit: 2.153) |
|---|---|---|---|
| 679 |  | 4.41 | 0.03 |
| 604 |  | 4.41 | 0.01 |
| 958 |  | 4.41 | 0.06 |
| 676 |  | 4.40 | 0.01 |
| 678 |  | 4.40 | 0.05 |
| 677 |  | 4.39 | 0.02 |
| 107 |  | 4.39 | 0.02 |
| 368 |  | 4.38 | 0.10 |
| 293 |  | 4.37 | 0.09 |
| 493 |  | 4.37 | 0.10 |

*(continued)*

**Table 4** (continued)

| ID | Compound | Predicted value<br>PaCa2 cellular uptake<br>(log10 [NP]/cell) | Domain of applicability<br>(limit: 2.153) |
|---|---|---|---|
| 494 |  | 4.37 | 0.10 |
| 550 |  | 4.36 | 0.11 |
| 196 |  | 4.35 | 0.06 |
| 200 |  | 4.35 | 0.06 |
| 626 |  | 4.35 | 0.06 |
| 602 |  | 4.35 | 0.06 |
| 981 |  | 4.34 | 0.10 |
| 925 |  | 4.34 | 0.10 |
| 192 |  | 4.34 | 0.06 |
| 65 |  | 4.34 | 0.05 |

In this study, the modeling and prediction of corrosion inhibition for steel in acidic environment through the development of QSPR with the aid of the Enalos KNIME nodes is described (Berthold et al. 2008; Melagraki and Afantitis 2013). The development of a predictive kNN model was realized by first calculating Mold2 molecular descriptors for the organic inhibitors with the Enalos Mold2 KNIME node. The predictive model was assessed with the Enalos Model Acceptability Criteria KNIME node. The domain of model applicability was determined using the Enalos Domain KNIME nodes.

Corrosion inhibition data for steel in acidic medium from various organic chemicals were collected from the literature (El Ashry et al. 2012) and were assembled in a single database. Inhibitors involve triazole, oxadiazole, and thiadiazole derivatives; aromatic hydrazides and Schiff bases; benzimidazole and 2- substituted derivatives; as well as pyridine derivatives. A set of 55 organic inhibitors in different concentrations yielded a total of 186 inhibition data.

The structural features of the studied corrosion inhibitors were evaluated with the Mold2 software. A total of 777 descriptors were initially calculated for each compound based on topological, geometrical, and structural criteria. From these, only 320 descriptors were used as possible inputs for the construction of the QSPR model, since the remaining descriptors were filtered out due to poor discrimination power (Ojha and Roy 2011).

Among the different modeling methodologies screened in KNIME platform, the k-nearest neighbors (kNN) technique (with an optimized value k = 3) was selected as the most appropriate for the specific data (Hall et al. 2009). Details on the kNN algorithm and methodology can be found elsewhere (Franco-Lopez et al. 2001).

The predictive model was validated internally and externally according to the standards of QSAR model acceptance, as imposed by the OECD. The complete dataset was randomly split into 70:30 ratio (training set: validation set) with the partitioning KNIME node. The combinations in the test set did not participate in the training procedure. The statistical criteria that were used to determine the robustness, reliability, and predictive ability of the model are: the coefficient of determination between experimental values and predicted values ($R^2$), validation via external test set, leave-one-out cross validation procedure and quality of fit and predictive ability of a continuous QSAR model based on Tropsha's tests (Tropsha 2010). The statistics of the validation procedure are shown in Table 5.

**Table 5** Statistical parameters of the QSPR model

| $R^2$ training (n = 131) | 0.96 |
|---|---|
| RMSE training | 4.90 |
| $R^2_{LOO}$ | 0.73 |
| $R^2$pred (n=55) | 0.84 |
| $RMSE_{pred}$ | 9.83 |

The predictive scheme included a KNIME workflow, which operated the following actions:

1. Compounds along with corrosion inhibition data were uploaded and preprocessed.
2. The calculation and selection of descriptors was realized.
3. The kNN methodology was carried out.
4. The produced model was validated.
5. The domain of applicability was determined.

The initial dataset of 186 corrosion inhibitors was randomly divided into 131 training set compounds and 55 validation set compounds (ratio 70:30). As mentioned above, only compounds from the training set were used to develop the QSPR models, and 320 descriptors were selected as possible inputs during the development.

The CfsSubset variable selection with the BestFirst evaluator method (Witten et al. 2005) was next applied on the training set to identify the most significant among the 320 descriptors. Thus, the concentration along with seven descriptors was selected as the most important parameter for the model development.

The selected descriptors include number of oxygen (D026), structural information content order-1 index (D282), Geary topological structure autocorrelation length-8 weighted by atomic Sanderson electronegativities (D470), Moran topological structure autocorrelation length-7 weighted by atomic polarizabilities (D509), lowest eigenvalue from Burden matrix weighted by van der Waals order-6 (D545), highest eigenvalue from Burden matrix weighted by van der Waals order-4 (D575), number of group Ar-CH=X (D741), and the concentration (C in mM).

The description of the selected descriptors will be briefly discussed below (Todeschini and Consonni 2009).

A combination of the descriptors offers a unified representation of the compound and high selective power. Descriptors D026 and D741 indicate the number of oxygen atoms and Ar-CH=X groups that may be present in the compound. Descriptor D282 encodes the structural information content order-1 index. This descriptor represents a graph theoretical invariant, which considers the molecular graph as being a source of different probability distributions, to which the information theory is applied (Todeschini and Consonni 2009). D470 encodes information as described by Geary topological structure autocorrelation length-8 weighted by atomic Sanderson electronegativities. The Geary index denotes spatial autocorrelation and is a distance-type function which varies from zero to infinite. D509 encodes information regarding the atomic polarizabilities combined with Moran topological structure autocorrelation length-7. Moran coefficient (value range between $-1$ and $+1$) indicates the spatial autocorrelation and is associated with atomic properties, the number of atoms, and the topological distance between them. Descriptors D545 and D575 represent the lowest eigenvalue from Burden matrix (Burden 1989) weighted by van der Waals order-6 and the highest eigenvalue from Burden matrix weighted by van der Waals order-4, respectively (Burden et al. 2009). Burden descriptors weighted by van der Waals have been shown to be very selective descriptors and in turn useful for similarity searching (Todeschini and Consonni 2009).

The above descriptors are associated with different weights that affect the corrosion inhibition across compounds. Therefore, according to the positive or negative impact of each descriptor, novel compounds with specified properties may be designed.

After comparison with the results from other methodologies (Melagraki and Afantitis 2013), it was concluded that the kNN approach yielded an accurate and powerful model that reliably predicts the efficiency of corrosion inhibition. Then, one may safely conclude that the selected descriptors encode the structural characteristics of the substances related to corrosion inhibition.

The experimental vs. predicted corrosion inhibition values for the training set and test set compounds are shown in Fig. 3. Outliers have been indicated and presented in the original article (Melagraki and Afantitis 2013).

The Enalos Model Acceptability Criteria KNIME node has been applied to the data (Fig. 4) and the model passed Tropsha's requirements for predictive ability.

$R^2$ is the determination coefficient between experimental and predicted values and model prediction on the test set ($R^2_{pred}$). The model was particularly stable with respect to the inclusion/exclusion of compounds measured by the leave-one-out (LOO) cross validation procedure ($R^2_{LOO} = 0.73$).

Another measure of robustness and statistical significance of a QSPR model is the Y-randomization test, which further validated our approach. An additional validation test was conducted to evaluate the predictive power of the method independently of the partitioning of the dataset (Melagraki and Afantitis 2013).

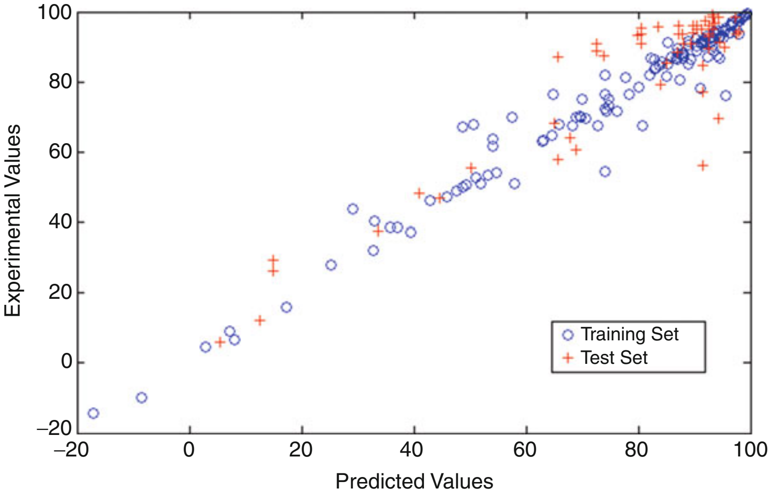
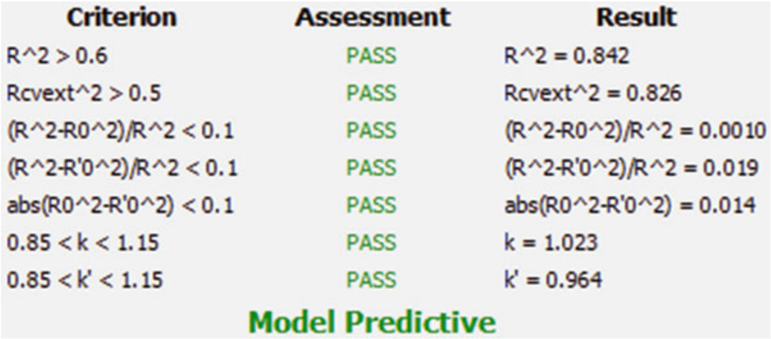


**Fig. 3** Experimental vs. predicted values for the training and test set



| Criterion | Assessment | Result |
|---|---|---|
| R^2 > 0.6 | PASS | R^2 = 0.842 |
| Rcvext^2 > 0.5 | PASS | Rcvext^2 = 0.826 |
| (R^2-R0^2)/R^2 < 0.1 | PASS | (R^2-R0^2)/R^2 = 0.0010 |
| (R^2-R'0^2)/R^2 < 0.1 | PASS | (R^2-R'0^2)/R^2 = 0.019 |
| abs(R0^2-R'0^2) < 0.1 | PASS | abs(R0^2-R'0^2) = 0.014 |
| 0.85 < k < 1.15 | PASS | k = 1.023 |
| 0.85 < k' < 1.15 | PASS | k' = 0.964 |

Model Predictive

**Fig. 4** Enalos Model Acceptability Criteria KNIME node screenshot

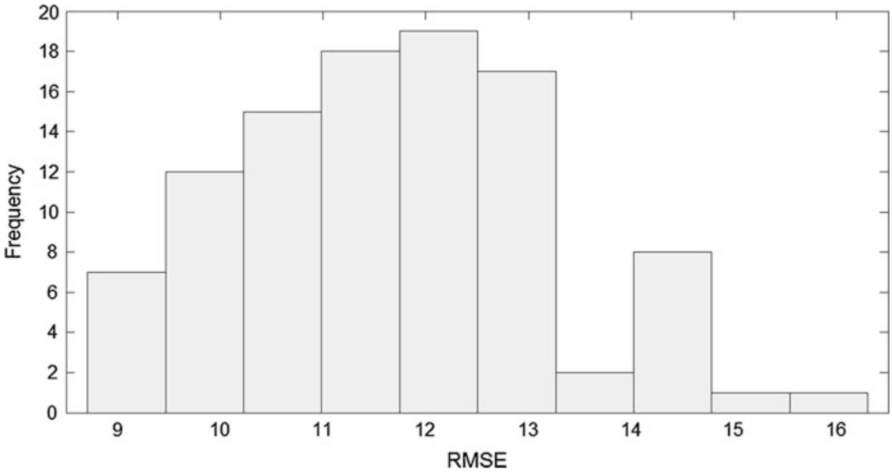


**Fig. 5** Distribution of the RMSE values (100 random splits)

The distribution of the root mean squared error (RMSE) values is shown in Fig. 5.

The applicability domain cutoff value was estimated to be 3.774 and 0.183 for similarity (Mouchlis et al. 2012) and leverage (Afantitis et al. 2008) calculations, respectively. Similarity calculations for all compounds in the test set had values which range between 0.015 and 1.23. However, the leverage predicted response of a simple pyridine (0.378) resulted from a significant model extrapolation, and it is the only prediction that may be considered unreliable.

The present approach, due to its high predictive power and the minimal requirement of only 2D structure information of a compound, could be a very useful tool for the determination of the corrosion inhibition. Moreover, this modeling method considerably decreases the time and cost required to experimentally design corrosion inhibitors. Also, the method may be applied to the screening of regular or virtual chemical databases, thus seeking new organic compounds with specific properties. For this purpose, the applicability domain will be an invaluable tool for discarding "divergent" chemical structures.

## KNIME-Based Design and Prediction of Compound Structures

The broad applicability of KNIME in the area of medicinal and pharmaceutical chemistry has also rendered it a valuable tool for structure prediction and analysis. Thus, KNIME workflows have been implemented over the years to perform drug design, virtual screening, molecular modeling, QSAR, QSPR, structure classification, and clustering applications among others.

In this section, we present two KNIME applications regarding the prediction of yellow fever inhibitors and a knowledge-based method for the de novo design of novel compounds.

### Prediction of Yellow Fever Inhibitors from ChEMBL Database Through KNIME Classification Analysis

Yellow fever (YF) is an acute infection, which is transmitted by arthropods and mosquitoes to humans (Agnihotri et al. 2012), and is caused by the mosquito-borne yellow fever flavivirus (YFV). The family of flaviviruses also includes other RNA viruses, such as the hepatitis C virus (HCV), the dengue virus (DENV), the West Nile virus (WNV), the Japanese encephalitis virus (JEV), and the bovine viral diarrhea virus (BVDV) among others (Agnihotri et al. 2012; Julander 2013).

YFV is a major health risk in particular regions of South America and Africa since almost 200,000 new infections and 30,000 deaths are observed every year (Chatelain et al. 2013); importantly, the fatality rate may reach 60 % in severe cases.

Currently, several regular antiviral drugs have been tested against YF disease, but no chemotherapeutic medication has been developed to specifically target YFV. An anti-YFV vaccine (17D) is used to prevent the infection; however, it has been observed to cause systemic infections and side effects in some patients (Julander 2013).

Therefore, computational approaches, such as virtual screening and modeling are particularly suitable to assist the discovery of new anti-YFV compounds. In this context, Moorthy and Poongavanam collected a number of compounds from the literature that were found (experimentally or computationally) to inhibit the YFV and employed them for the development of KNIME classification models using the Naive Bayes approach (Narayana Moorthy and Poongavanam 2015).

For this purpose, a set of 379 YFV inhibitors were collected from the ChEMBL database (https://www.ebi.ac.uk/chembl/). After the initial treatment of the data (namely, salt removal, 3D structure generation, and energy minimization of the structures), 30 two-dimensional descriptors of the compounds were calculated with the CDK tool as implemented in KNIME (Berthold et al. 2008; Beisken et al. 2013). In total, 16 classification models were developed using the Weka data mining software (Hall et al. 2009). The first 12 models refer to individual datasets, while models 13–16 were developed from a combined dataset, which contained 309 compounds. Before the development of the models, the dataset was partitioned into training (65 %) and test (35 %) set based on a sampling consideration that distributes the inhibitors and noninhibitors evenly between sets. The inhibitor definition was restricted to specific activity criteria (i.e., IC50 $\leq$ 10 μM for inhibitors; IC50 $>$ 10 μM for noninhibitors). The quality of models was investigated through various activity thresholds (10, 30, 50, and 100 μM) and statistical parameters, such as the sensitivity, specificity, G-mean, Matthew's correlation coefficient (MCC), and overall accuracy.

Principal component analysis (PCA) revealed that the dataset does not contain distinct clusters and its diversity is adequately represented by the training set. Despite the presence of some outliers, it was observed that they were not structurally similar whatsoever. Additionally, it was shown that the majority of inhibitors is highly affected by the topological polar surface area (TPSA) and polar bonds. This indicated the more hydrophobic nature of noninhibitors compared to inhibitors.

The BestFirst attribute selection module of Weka was employed to select a set of 24 physicochemical descriptors in order to construct the Naive Bayes classification models for YFV inhibition. It was shown that the six datasets (activity cutoff 30 μM) perform equally well with an overall accuracy for the test set being >75 %.

The model quality (based on MCC) was higher for all datasets except for one, which performed poorly. Models derived from all datasets were statistically significant with MCC and G-mean values >0.7. The significance of the dataset was also verified by the high F–score values (>0.8), while sensitivity and specificity parameters were >0.75 for most models. The F–score is calculated through the following equation:

$$\text{F–score} = \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \qquad (1)$$

Also, the predictive ability of the scheme was investigated by combining all the datasets into one. Thus, 70 compounds (Krečmerová et al. 2007) were used in the test set, while the remaining 309 compounds were combined into one dataset to train the model. It was observed that an activity cutoff of 50 μM yields a relatively good accuracy. Despite that models developed by different activity thresholds (10 or 30 μM) were successful in distinguishing inhibitors (>90 %) from noninhibitors ( 65 %), they lacked a balanced class distribution (predicting adequately both classes instead of predicting accurately only one), which clearly affects the quality of the model. Therefore, new models were constructed based on the 309 compounds set (200 compounds in the training set, 109 compounds in the test set). From the models for each activity threshold (10, 30, 50, and 100 μM), one model (at 50 μM) displayed superior performance over the others. That is, it predicted 92 % of inhibitors and 78 % of noninhibitors with a MCC coefficient of 0.71, and a G-mean score (0.84), which reflects the high quality. Finally, the best YFV inhibition model (at 50 μM) was implemented into KNIME workflow to become freely available for use in medicinal chemistry applications.

## De Novo Design of Synthetically Feasible Compounds Through the Reaction Vectors Approach

Early attempts to develop successful de novo design programs for compounds have been seriously hampered by the unfeasible synthesis of the proposed structures (Boda and Johnson 2006; Lewis and Leach 1994; Schneider and Fechner 2005; Gillet and Johnson 1998). More recent approaches that may facilitate synthetic routes generate a fixed set of transformations, which are in turn applied to starting structures to generate new molecules (Fechner and Schneider 2006, 2007; Lameijer et al. 2006; Schürer et al. 2005; Vinkers et al. 2003). However, these methods are usually restricted by the limited number of "reactions" that can be performed.

Reaction-like methodologies have been used in QSAR and data mining studies. For instance, the "matched molecular pairs" scheme (Leach et al. 2006) evaluates the change of a property with respect to a single structural change. Similar approaches perform clustering of molecule pairs according to descriptor difference vectors, which are constructed by subtracting the vector representation of the "product" molecule from the vector representation of the "reactant" molecule (Sheridan et al. 2006). Therefore, the pairs of molecules which belong to the same cluster represent similar transformations. Thus, changes in activities associated with the pairs may be used to estimate the effect of a particular transformation on activity.

A knowledge-based approach regarding the de novo design of compounds based on reaction vectors has been proposed by Patel et al. (Patel et al. 2009). The method of reaction vectors is related to the descriptor difference vectors approach (Sheridan et al. 2006) and characterizes the structural changes, which occur at the reaction center.

The authors' flexible procedure enabled the automatic collection of ever-increasing data on reactions, which are available in several databases. Reaction vectors could be applied to novel starting molecules through a structure generation algorithm to predict new structures. Each applied transformation is derived from a reactions database, and since it is associated with a reaction from the literature, one may be confident toward the synthetic feasibility of the proposed molecules. The vectors are automatically collected from a set of reactions, which is not restricted by size or reaction type, therefore without involving any complex reaction strategy.

The principle of the method is the utilization of the information taken from a set of reactions (parent reactions) in deriving reaction vectors. Reaction vectors are used to describe the reaction environment as well as changes at the reaction center. Then, the reaction vectors are extracted from the knowledge base and may be applied to input starting materials in order to predict new product compounds for synthesis. In some cases, successful predictions may require the operation of a reaction vector to a second starting molecule.

In the past years, reaction vectors have been used to search and classify chemical reactions (Broughton et al. 2003). As already mentioned, a reaction vector represents the changes that occur during a reaction. This is described in a descriptor scheme that is enhanced in the products (positive descriptors, they represent bonds gained in products) and is diminished from the reactants (negative descriptors, they represent bonds lost from reactants). In de novo design, the practical role of reaction vectors is to propose synthetic ways to new compounds by achieving a balance between specificity and generality. Namely, the reaction vectors should be specific enough in order to avoid their application in environments which disfavor the course of a reaction, but encoding part of the environment will hamper their ability to generate novel structures. It was shown that such a balance may be obtained by combining atom pairs separated by one and two bonds (Patel et al. 2007). The reaction vectors employed by Patel et al. are modifications of the original descriptors developed by Carhart et al. (1985). More detailed information on reaction vectors can be found in Patel et al. and references therein (Patel et al. 2009).

The authors revealed that many reactions in the Lilly database are incomplete, and they implemented a reaction cleaning algorithm to overcome this drawback. The reaction cleaning algorithm is applied to the reactions before calculating the atom pair descriptors for the reactants and products. The reaction vector is next calculated as follows:

$$D = P\text{–}R \tag{2}$$

where D is the reaction vector, P is the product vector (the sum of the vectors of individual products), and R is the reactant vector (the sum of the vectors of individual reactants). The algorithms have been implemented with the aid of the JoeLib toolkit (joelib.sourceforge.net/) with the de novo design tool available in KNIME.

The de novo design algorithm was validated internally by reproducing reactions in the Lilly database and externally by reproducing the already known synthetic pathways of two drugs. Internal validation was achieved by generating 90 % of the known parent products for each reaction in a dataset containing 5695 reactions and 2866 reaction vectors. The external validation involved the syntheses of the intermediate of the antithrombotic drug (S)-(+)-clopidogrel bisulfate (Wang et al. 2007) and the antidepressant venlafaxine (Kavitha and Rangappa 2004). In each case, the products were successfully generated using a 5839 reagents set and a dataset, which comprised 24,418 reactions and 16,859 reaction vectors. The reaction vectors procedure was successful in reproducing the known synthetic routes for both drugs; however, it was suggested that this approach also yields quite a few alternative products.

The applicability of the algorithm to de novo design was demonstrated in three cases: (1) the ability of a reaction vector to generate novel and diverse products as regards the parent reaction, (2) prediction of analogs of a lead compound, and (3) application to the enumeration of a compound library.

In the first case, 10 reactants were randomly selected from the smaller (2866 vectors/5695 reactions) dataset and were input to the de novo design tool. Products were generated employing the 16,859 reaction vectors from the larger dataset. Structures containing more functional groups than others usually yielded more solutions, therefore, the number of products that were created lies within 0 and 44. Moreover, the average similarities between predicted compounds and products of the parent reactions vary significantly (0.15–0.96). It is therefore suggested that the diversity of the products depends on both the starting molecule and the entries in the reaction database.

Next, the predictive power of the algorithm in lead optimization cases was explored by considering drugs as starting compounds and generating products after single-step transformations. The products were assumed to be synthetically feasible if they were identified upon search in SciFinder Scholar. The procedure was based on penicillin G, Prozac, and aspirin (starting structures) and was performed with the larger set of reaction vectors (16,859) and the 5839 reagent set as described above. Penicillin G was associated with the most potential products (24), while 20 and 12 products were generated for Prozac and aspirin, respectively. The reaction vector methodology was shown to realize the generation of synthetically accessible molecules that share close structural similarities with a parent compound. Thus, this approach could be particularly valuable in structure-activity relationship (SAR) studies by proposing structures for synthesis that explore sufficiently the conformational space around a known lead compound.

Finally, the proposed methodology can be used to construct a virtual library of products by listing all possible products from a single reaction employing a specific set of reagents. For instance, the authors used 6-bromoquinoxalin-2-one as starting material, 628 boronic acids as the reagents (extracted from the ACD), and a Suzuki coupling reaction to construct an enumerated library with 292 compounds. It was shown that this scheme may be used to direct a parallel synthesis route according to a single reaction with varying reagents.

Moreover, the way the reaction vector methodology recognizes the environment of a reaction is clearly demonstrated. For the transformation that is applied, there is a literature precedence, and thus a high level of confidence is achieved regarding the completion of the reaction. However, this requires that complete coverage of a generic reaction has been obtained through adequate examples in the reaction database.

The approach has been implemented in KNIME and this allows the facile customization of the procedure for a variety of applications. Overall, the usefulness of the proposed algorithm is to provide sets of molecules, which are associated with multiple purposes, and at the same time being synthetically feasible.

**Acknowledgments** This work is supported by funding under the Seven Research Framework Programme of the European Union: Project THALAMOSS (HEALTH.2012.1.2-1 grant agreement no. 306201) and Project nanoMILE (grant agreement no. 310451).

## Bibliography

Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P. A., Markopoulos, J., & Igglessi-Markopoulou, O. (2008). Development and evaluation of a QSPR model for the prediction of diamagnetic susceptibility. *QSAR & Combinatorial Science, 27*(4), 432–436.

Afantitis, A., Melagraki, G., Koutentis, P. A., Sarimveis, H., & Kollias, G. (2011). Ligand – based virtual screening procedure for the prediction and the identification of novel β-amyloid aggregation inhibitors using Kohonen Maps and Counterpropagation Artificial Neural Networks. *European Journal of Medicinal Chemistry, 46*, 497–508.

Agnihotri, S., Narula, R., Joshi, K., Rana, S., & Singh, M. (2012). In silico modeling of ligand molecule for non structural 3 (NS3) protein target of flaviviruses. *Bioinformation, 8*(3), 123–127.

Beisken, S., Meinl, T., Wiswedel, B., de Figueiredo, L. F., Berthold, M., & Steinbeck, C. (2013). KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinformatics, 14*, 257–257.

Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2008). KNIME: The Konstanz information miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 319–326). Berlin/Heidelberg: Springer.

Boda, K., & Johnson, A. P. (2006). Molecular complexity analysis of de novo designed ligands. *Journal of Medicinal Chemistry, 49*(20), 5869–5879.

Broughton, H., Hunt, P., & MacKey, M. (2003) Methods for classifying and searching chemical reactions. Google Patents.

Brown, F. K. (1998). Chemoinformatics, what it is and how does it impact drug discovery. *Annual Reports in Medicinal Chemistry, 33*, 375–384.

Burden, F. (1989). Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences, 29*, 225–227.

Burden, F., Polley, M., & Winkler, D. (2009). Toward novel universal descriptors: Charge fingerprints. *Journal of Chemical Information and Modeling, 49*, 710–715.

Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: Definition and applications. *Journal of Chemical Information and Computer Sciences, 25*(2), 64–73.

Chatelain, G., Debing, Y., Burghgraeve, T. D., Zmurko, J., Saudi, M., Rozenski, J., Neyts, J., & Aerschot, A. V. (2013). In search of flavivirus inhibitors: Evaluation of different tritylated nucleoside analogues. *European Journal of Medicinal Chemistry, 65*, 249–255.

Cohen, Y., Rallo, R., Liu, R., & Liu, H. H. (2013). In silico analysis of nanomaterials hazard and risk. *Accounts of Chemical Research, 46*(3), 802–812.

Ebenso, E. E., Kabanda, M. M., Murulana, L. C., Singh, A. K., & Shukla, S. K. (2012). Electrochemical and quantum chemical investigation of some azine and thiazine dyes as potential corrosion inhibitors for mild steel in hydrochloric acid solution. *Industrial and Engineering Chemistry Research, 51*, 12940–12958.

El Ashry, E. S. H., El Nemr, A., & Ragab, S. (2012). Quantitative structure activity relationships of some pyridine derivatives as corrosion inhibitors of steel in acidic medium. *Journal of Molecular Modeling, 18*, 1173–1188.

Epa, V. C., Burden, F. R., Tassa, C., Weissleder, R., Shaw, S., & Winkler, D. A. (2012). Modeling biological activities of nanoparticles. *Nano Letters, 12*(11), 5808–5812.

Fechner, U., & Schneider, G. (2006). Flux (1): A virtual synthesis scheme for fragment-based de novo design. *Journal of Chemical Information and Modeling, 46*(2), 699–707.

Fechner, U., & Schneider, G. (2007). Flux (2): Comparison of molecular mutation and crossover operators for ligand-based de novo design. *Journal of Chemical Information and Modeling, 47*(2), 656–667.

Fourches, D., Pu, D., Tassa, C., Weissleder, R., Shaw, S. Y., Mumper, R. J., & Tropsha, A. (2010). Quantitative nanostructure−activity relationship modeling. *ACS Nano, 4*(10), 5703–5712.

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment, 77*, 251–274.

Gajewicz, A., Rasulev, B., Dinadayalane, T. C., Urbaszek, P., Puzyn, T., Leszczynska, D., & Leszczynski, J. (2012). Advancing risk assessment of engineered nanomaterials: Application of computational approaches. *Advanced Drug Delivery Reviews, 64*(15), 1663–1693.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research, 40*(Database issue), D1100–D1107.

Gillet, V. J., & Johnson, A. P. (1998). *Structure generation for De Novo design*. Washington: American Chemical Society.

Gutlein, M., Karwath, A., & Kramer, S. (2012). CheS-mapper – chemical space mapping and visualization in 3D. *Journal of Cheminformatics, 4*(1), 7.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, 11*(1), 10–18.

Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., Su, Z., Perkins, R., & Tong, W. (2008). Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of Chemical Information and Modeling, 48*, 1337–1344.

Hu, Y., & Bajorath, J. (2012). Freely available compound data sets and software tools for chemoinformatics and computational medicinal chemistry applications [version 1; referees: 2 approved]. Vol. 1.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., & Coleman, R. G. (2012). ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling, 52*(7), 1757–1768.

Julander, J. G. (2013). Experimental therapies for yellow fever. *Antiviral Research, 97*(2), 169–179.

Kavitha, B. C. V., & Rangappa, K. S. (2004). Simple and an efficient method for the synthesis of 1-[2-dimethylamino-1-(4-methoxy-phenyl)-ethyl]-cyclohexanol hydrochloride: (±) venlafaxine racemic mixtures. *Bioorganic & Medicinal Chemistry Letters, 14*(12), 3279–3281.

Kleandrova, V. V., Luan, F., Gonzalez-Diaz, H., Ruso, J. M., Melo, A., Speck-Planche, A., & Cordeiro, M. N. (2014a). Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environment International, 73*, 288–294.

Kleandrova, V. V., Luan, F., Gonzalez-Diaz, H., Ruso, J. M., Speck-Planche, A., & Cordeiro, M. N. (2014b). Computational tool for risk assessment of nanomaterials: novel QSTR-perturbation model for simultaneous prediction of ecotoxicity and cytotoxicity of uncoated and coated nanoparticles under multiple experimental conditions. *Environmental Science and Technology, 48*(24), 14686–14694.

Krečmerová, M., Holý, A., Pískala, A., Masojídková, M., Andrei, G., Naesens, L., Neyts, J., Balzarini, J., De Clercq, E., & Snoeck, R. (2007). Antiviral Activity of Triazine Analogues of 1-(S)-[3-Hydroxy-2-(phosphonomethoxy)propyl]cytosine (Cidofovir) and related compounds. *Journal of Medicinal Chemistry, 50*(5), 1069–1077.

Lameijer, E.-W., Kok, J. N., Bäck, T., & Ijzerman, A. P. (2006). The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *Journal of Chemical Information and Modeling, 46*(2), 545–552.

Leach, A. R., & Gillet, V. J. (2007). *An introduction to chemoinformatics* (Rev. ed.). Dordrecht: Springer.

Leach, A. G., Jones, H. D., Cosgrove, D. A., Kenny, P. W., Ruston, L., MacFaul, P., Wood, J. M., Colclough, N., & Law, B. (2006). Matched molecular pairs as a guide in the optimization of pharmaceutical properties; A study of aqueous solubility, plasma protein binding and oral exposure. *Journal of Medicinal Chemistry, 49*(23), 6672–6682.

Lee, A., Mercader, A. G., Duchowicz, P. R., Castro, E. A., & Pomilio, A. B. (2012). QSAR study of the DPPH radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes, halophenols and caffeic acid analogues. *Chemometrics and Intelligent Laboratory Systems, 116*, 33–40.

Lewis, R., & Leach, A. (1994). Current methods for site-directed structure generation. *Journal of Computer-Aided Molecular Design, 8*(4), 467–475.

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research, 35*(Database issue), D198–D201.

Liu, R., Rallo, R., George, S., Ji, Z., Nair, S., Nel, A. E., & Cohen, Y. (2011). Classification nano-SAR development for cytotoxicity of metal oxide nanoparticles. *Small, 7*(8), 1118–1126.

Liu, R., Rallo, R., Weissleder, R., Tassa, C., Shaw, S., & Cohen, Y. (2013). Nano-SAR development for bioactivity of nanoparticles with considerations of decision boundaries. *Small, 9*(9–10), 1842–1852.

Melagraki, G., & Afantitis, A. (2013). Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium. *Chemometrics and Intelligent Laboratory Systems, 123*, 9–14.

Melagraki, G., & Afantitis, A. (2014). Enalos InSilicoNano platform: An online decision support tool for the design and virtual screening of nanoparticles. *RSC Advances, 4*, 50713–50725.

Melagraki, G., & Afantitis, A. (2015). A risk assessment tool for the virtual screening of metal oxide nanoparticles through enalos insiliconano platform. *Current Topics in Medicinal Chemistry, 15*(18), 1827–1836.

Mouchlis, V. D., Melagraki, G., Mavromoustakos, T., Kollias, G., & Afantitis, A. (2012). Molecular modeling on pyrimidine-urea inhibitors of TNF-α production: An integrated approach using a combination of molecular docking, classification techniques, and 3D-QSAR CoMSIA. *Journal of Chemical Information and Modeling, 52*, 711–723.

Narayana Moorthy, N. S. H., & Poongavanam, V. (2015). The KNIME based classification models for yellow fever virus inhibition. *RSC Advances, 5*(19), 14663–14669.

O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., & Hutchison, G. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics, 3*(1), 33.

Ojha, P. K., & Roy, K. (2011). Comparative QSARs for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection. *Chemometrics and Intelligent Laboratory Systems, 109*(2), 146–161.

Papa, E., Kovarich, S., & Gramatica, P. (2009). Development, validation and inspection of the applicability domain of QSPR models for physicochemical properties of polybrominated diphenyl ethers. *QSAR Combinatorial Science, 28*, 790–796.

Patel, H., Gillet V. J., Chen, B., & Bodkin, M. J. (2007). *Development of a de novo design tool using reaction vectors*. In Poster presented at the 4th Joint Sheffield Conference on Chemoinformatics Sheffield, UK.

Patel, H., Bodkin, M. J., Chen, B., & Gillet, V. J. (2009). Knowledge-based approach to de novo design using reaction vectors. *Journal of Chemical Information and Modeling, 49*(5), 1163–1184.

Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Michalkova, A., Hwang, H.-M., Toropov, A., Leszczynska, D., & Leszczynski, J. (2011). Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature Nanotechnology, 6*(3), 175–178.

Russo, E. (2002). Chemistry plans a structural overhaul. *Nature, 419*(6903), 4–7.

Schneider, G., & Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nature Reviews. Drug Discovery, 4*(8), 649–663.

Schürer, S. C., Tyagi, P., & Muskal, S. M. (2005). Prospective exploration of synthetically feasible, medicinally relevant chemical space. *Journal of Chemical Information and Modeling, 45*(2), 239–248.

Shao, C. Y., Chen, S. Z., Su, B. H., Tseng, Y. J., Esposito, E. X., & Hopfinger, A. J. (2013). Dependence of QSAR models on the selection of trial descriptor sets: A demonstration using nanotoxicity endpoints of decorated nanotubes. *Journal of Chemical Information and Modeling, 53*(1), 142–158.

Shaw, S. Y., Westly, E. C., Pittet, M. J., Subramanian, A., Schreiber, S. L., & Weissleder, R. (2008). Perturbational profiling of nanomaterial biologic activity. *Proceedings of the National Academy of Sciences of the United States of America, 105*(21), 7387–7392.

Sheridan, R. P., Hunt, P., & Culberson, J. C. (2006). Molecular transformations as a way of finding and exploiting consistent local QSAR. *Journal of Chemical Information and Modeling, 46*(1), 180–192.

Speck-Planche, A., Kleandrova, V. V., Luan, F., & Cordeiro, M. N. (2015). Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine (London, England), 10*(2), 193–204.

Steinbeck, C., Han, Y. Q., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. L. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences, 43*(2), 493–500.

Sushko, I., Novotarskyi, S., Körner, R., Pandey, A. K., Rupp, M., Teetz, W., Brandmaier, S., Abdelaziz, A., Prokopenko, V. V., Tanchuk, V. Y., Todeschini, R., Varnek, A., Marcou, G., Ertl, P., Potemkin, V., Grishina, M., Gasteiger, J., Schwab, C., Baskin, I. I., Palyulin, V. A., Radchenko, E. V., Welsh, W. J., Kholodovych, V., Chekmarev, D., Cherkasov, A., Aires-de-Sousa, J., Zhang, Q.-Y., Bender, A., Nigsch, F., Patiny, L., Williams, A., Tkachenko, V., & Tetko, I. V. (2011). Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. *Journal of Computer-Aided Molecular Design, 25*(6), 533–554.

Tetko, I. V. (2012). The perspectives of computational chemistry modeling. *Journal of Computer-Aided Molecular Design, 26*, 135–136.

Todeschini, R., & Consonni, V. (2009). *Molecular descriptors for chemoinformatics*. Weinheim: Wiley.

Toropov, A. A., Toropov, A. P., Martyanov, S. E., Benfenati, E., Gini, G., Leszczynska, D., & Leszczynski, J. (2012). CORAL: Predictions of rate constants of hydroxyl radical reaction using representation of the molecular structure obtained by combination of SMILES and Graph approaches. *Chemometrics and Intelligent Laboratory Systems, 112*, 65–70.

Toropov, A. A., Toropova, A. P., Puzyn, T., Benfenati, E., Gini, G., Leszczynska, D., & Leszczynski, J. (2013). QSAR as a random event: Modeling of nanoparticles uptake in PaCa2 cancer cells. *Chemosphere, 92*(1), 31–37.

Tropsha, A. (2010). Best Practices for QSAR model development, validation, and exploitation. *Molecular Informatics, 29*(6–7), 476–488.

Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F. D., Heeres, J., Koymans, L. M. H., van Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., & Janssen, P. A. J. (2003). Synopsis: Synthesize and optimize system in Silico. *Journal of Medicinal Chemistry, 46*(13), 2765–2773.

Vrontaki, E., Mavromoustakos, T., Melagraki, G., & Afantitis, A. (2015). Quantitative nanostructure-activity relationship models for the risk assessment of nanomaterials. In K. Roy (Ed.), *Quantitative structure-activity relationships in drug design, predictive toxicology, and risk assessment* (pp. 537–561). Hershey, PA: IGI Global.

Wang, L., Shen, J., Tang, Y., Chen, Y., Wang, W., Cai, Z., & Du, Z. (2007). Synthetic improvements in the preparation of clopidogrel. *Organic Process Research & Development, 11*(3), 487–489.

Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., & Bryant, S. H. (2009). PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research, 37*(Web Server issue), W623–W633.

Weissleder, R., Kelly, K., Sun, E. Y., Shtatland, T., & Josephson, L. (2005). Cell-specifc targeting of nanoparticles by multivalent attachment of small molecules. *Nature Biotechnology, 23*, 1418–1423.

Winkler, D. A., Mombelli, E., Pietroiusti, A., Tran, L., Worth, A., Fadeel, B., & McCall, M. J. (2013). Applying quantitative structure-activity relationship approaches to nanotoxicology: Current status and future potential. *Toxicology, 313*(1), 15–23.

Winkler, D. A., Burden, F. R., Yan, B., Weissleder, R., Tassa, C., Shaw, S., & Epa, V. C. (2014). Modelling and predicting the biological effects of nanomaterials. *SAR and QSAR in Environmental Research, 25*(2), 161–172.

Witten, I. H., Frank, E., & Hall, M. A. (2005). *Data mining, practical machine learning tools and techniques*. San Francisco, CA: Elsevier.

Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., & Tropsha, A. (1995). Novel Automated Lazy Learning QSAR (ALL-QSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALLQSAR models. *Journal of Chemical Information and Modeling, 46*, 1984–1995.

Zhang, H., Ji, Z., Xia, T., Meng, H., Low-Kam, C., Liu, R., Pokhrel, S., Lin, S., Wang, X., Liao, Y.-P., Wang, M., Li, L., Rallo, R., Damoiseaux, R., Telesca, D., Mädler, L., Cohen, Y., Zink, J. I., & Nel, A. E. (2012). Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation. *ACS Nano, 6*(5), 4349–4368.