A Platinum Open Access Journal for Organic Chemistry

Review

Free to Authors and Readers | DOAJ Seal | *Arkivoc* **2024** (1) 202412195

Application of artificial intelligence and machine learning methods in drug discovery and development

Carlos Naranjo-Castañeda,[a] Carlos A. Coello-Coello,[b,c] and Eusebio Juaristi[a,c]

[a] *Departamento de Química, Centro de Investigación y de Estudios Avanzados, Avenida IPN # 2508, Col. San Pedro Zacatenco, 07360 México, CDMX, México*

[b] *Departamento de Computación, Centro de Investigación y de Estudios Avanzados, Avenida IPN # 2508, Col. San Pedro Zacatenco, 07360 CDMX, México*

[c] *El Colegio Nacional, Donceles # 104, Centro Histórico, 06000 CDMX, México*

Emails: carlos.naranjo@cinvestav.mx, ccoello@cs.cinvestav.mx, ejuarist@cinvestav.mx

Dedicated to Professors Alan R. Katritzky and Charles W. Rees, in celebration of the 25th Anniversary of Arkivoc

Received 03-08-2024 **Accepted** 05-21-2024 **Published on line** 06-06-2024

Abstract

It is clear that the COVID-19 pandemic has brought renewed attention to the urgent need for the development of efficient strategies for drug discovery and development. In particular, there is an increasing demand for more powerful and reliable computational methods that contribute to satisfy needs in the health sector owing to the urgent need for effective pharmacological therapies. Here, we review *Artificial Intelligence* (AI) approaches for predicting biological activity and chemical properties of designed molecules, as well as for the planning and execution of experiments testing their pharmacological behavior. This review begins with a brief introduction to *Machine Learning* (ML) methods. Next, the virtual screening protocols that are commonly used in the domain of protein-ligand interactions, ligand binding affinity, and binding pose (conformation) are reviewed, including classical ML algorithms and deep learning methods. We also discuss the ML approach implemented to predict and design synthetic pathways to reach a molecule. Finally, the application in self-driving labs (SDLs) for the execution of experiments in organic synthesis is presented. It is hoped that this review will promote the exploration of more accurate ML-based prediction strategies to examine molecules with potential biological activity.

Keywords: Artificial intelligence, machine learning, drug discovery, self-driving laboratories

Cite as *Arkivoc* **2024** (1) 202412195
DOI: https://doi.org/10.24820/ark.5550190.p012.195

Table of Contents

1. Introduction

The recent COVID-19 pandemic has challenged the healthcare sector and has induced the pharmaceutical industry to undertake unprecedented scientific efforts to obtain an effective vaccine in the shortest possible time. Currently, the pharmaceutical industry is at an inflection point driven by the demand for faster, better, and cheaper drug development techniques to meet demand. In this regard, Artificial intelligence (AI) is being leveraged to accelerate drug discovery, which will eventually reduce the cost of treatments.[1-3] AI techniques such as machine learning (ML) and deep learning (DP) accelerate and improve the drug development process by enabling more efficient and accurate analysis of substantial amounts of data, such as building models to estimate and/or classify the bioactivity of new ligands, prediction of target structures, the optimization or discovery of 'hits' and 'hit' candidates ('hit' is a compound that exhibits the desired activity, which is confirmed upon reiterative testing[4]), as well as the elaboration of models that help predict pharmacokinetic and toxicological aspects (ADMET) of interest.[1-5]

The AI drug development process, driven by the availability of massive amounts of data and algorithms that can process them, enabled by enormous advances in computational processing, as well as a growing internet infrastructure, gave rise to an area called Big Data. In particular, big data contains information related to the pharmaceutical chemistry, which has the necessary components for the exploration of therapeutic targets (a pharmacological target or molecular target is defined as the place in the organism where a drug employs its action) and molecules with possible pharmacological activity.[6] In addition, information is available on molecular structure, clinical trial reports, patents, drugs on the market, as well as molecular building blocks or molecules that have been characterized by having some biological effect or activity.[7-9]

AI-assisted drug development begins with data preparation, in a procedure that goes from data accumulation, pre-processing, and transformation with the aim of translating chemical-pharmaceutical data into machine-readable representations. Of the protocols with ML in pharmaceutical development, one of the most common is "smart" virtual screening, which consists of the application of ML algorithms in the exploration of different databases of compounds or molecular fragments, in order to identify and select a number of compounds that present the desired biological activity on a specific therapeutic target. Another is computer-aided drug design that seeks to predict the best route for the synthesis of compounds of interest, based on criteria that evaluate the viability of the synthetic route and the investment costs for such synthesis. On the other hand, in self-driving labs (SDLs), algorithms are applied for decision-making in the sequencing and optimization of reaction conditions, as well as specialized software and robots to carry out the process, See Figure 1.

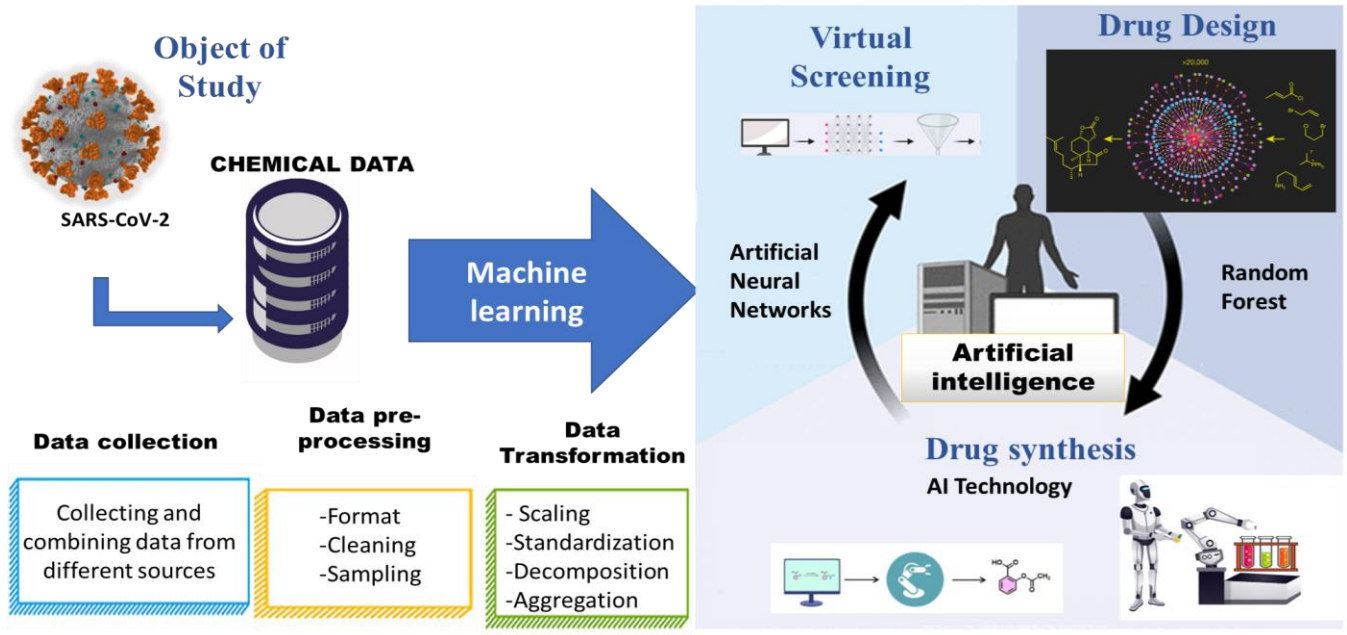

Figure 1. Data Preparation and AI-Assisted Drug Development Process.

2. Data Preparation

Before starting the construction of an ML protocol, it is necessary to collect data from the various sources of information and order it by its relevance. Ideally, such a dataset should contain structurally diverse molecules and should cover a wide range of values of the target property.[10-11] These data should contain information related to physicochemical properties (molecular weight, melting point), reaction conditions (temperature, pressure, and time), and fingerprints or segments that describe the structural information of the molecules.[12] Since the collected data may be in an unwanted, disorganized, or extremely large format, pre-processing is necessary to improve its quality. The three common steps for data preprocessing are *formatting*, *cleaning*, and *sampling*.[13] *Format* ensures that all variables within the same attribute are written consistently, while data *cleaning* involves removing messy data and incorporating missing values, whereas *sampling* is applied when too much data may result in impractical analysis. A further step is the transformation of the data, as most datasets contain features that vary in terms of range, units, or magnitude. Data needs to be transformed with *scaling*

and *standardization* in mind. Also, if some values in the dataset are too complex, *decomposition* into multiple constituent parts may be more meaningful to an ML model (an example is time constituted by date and time; in some cases, it is better to separate them and keep only the relevant data). The opposite situation may be observed when *aggregation of* related data (an example is toxicity and half-life, being related pharmacokinetic parameters) is useful for ML algorithms.

2.1. Learning algorithms

Depending on the data available and the task to perform, we can choose among several types of learning algorithms. The most common are 'supervised learning', 'unsupervised learning', 'semi-supervised learning' and 'reinforcement learning'.

In *supervised learning*, a function is deduced from a set of training data that is categorized or labeled using regression and classification. A regression task consists of predicting an objective numerical value, such as the conductivity, product yield, and adsorption capacity of the molecules when given a set of inputs, whereas a classification task pertains the selection of models based on input parameters and their corresponding output. This way, the algorithm learns to select the most feasible outcomes based on the correlation between the input data and the learned data. Frequently used supervised models include Artificial Neural Networks (ANNs), Random Forest (also known as random decision forest, RF), Support Vector Machines (SVMs), etc.[14] This methodology can be applied, for example, in Quantitative Relationship Structure Activity or Quantitative Relationship Structure Property (QSAR/QSRP) models.

Unsupervised learning employs unclassified or unlabeled datasets and allows the model to learn without any guidance. It consists of two common methods; that is, data grouping or dimensionality reduction. Data grouping involves initial calculation of the similarities of all samples according to specific metrics followed by assignment to specific groups based on their similarities, patterns, and dissimilarities. On the other hand, dimensionality reduction involves mapping a high-dimensional data matrix to a low-dimensional one while maintaining the information provided in the original data.[12] Basically, models must discover the hidden pattern within the unlabeled data, and then generate clusters of them. Artificial Neural Networks (ANNs)[15] derivations are included in this model, and can be used, for example, to find hidden patterns in medical and/or biological data, and to identify new drug targets relevant to the cure of diseases.[16,17]

Semi-supervised learning falls in between supervised learning and unsupervised learning. These methods have been effective when addressing the handling of incomplete databases, with the use of previously established criteria.[18,19]

Reinforcement learning examines the environment repetitively before taking action. This methodology aims to use experiences, which would either minimize risks or maximize benefits. The most commonly used algorithms are deep adversarial networks,[20] Time Difference[21] and Q-Learning.[22]

The learning models mentioned above must be evaluated to determine their performance. In practice, several model validation strategies can be used to prevent overfitting (overfitting is an unwanted ML behavior that occurs when the model provides accurate predictions for training data, but not for new data). In an ideal situation, the available data would be divided into 3 parts: training, validation, and test datasets.[23] There are many methods available for validating machine learning models, such as historical data validation, sensitivity analysis, predictive validation, comparison with other models, residual time evolution, Wilcoxon signed rank test, McNemar test, etc.[24,25]

Currently, the ML algorithms most widely used in the field of drug design are: *Random Forest* (RF) and *Artificial Neural Networks* (ANNs) and Support Vector Machines (SVMs) among their derivations for increased

accuracy and satisfactory performance on substantial amounts of data.[26] Other available algorithms are *k-nearest neighbors* (KNN),[27] *naïve Bayes* (NB),[28] and *logistic regression* (LR),[29] (Figure 2).

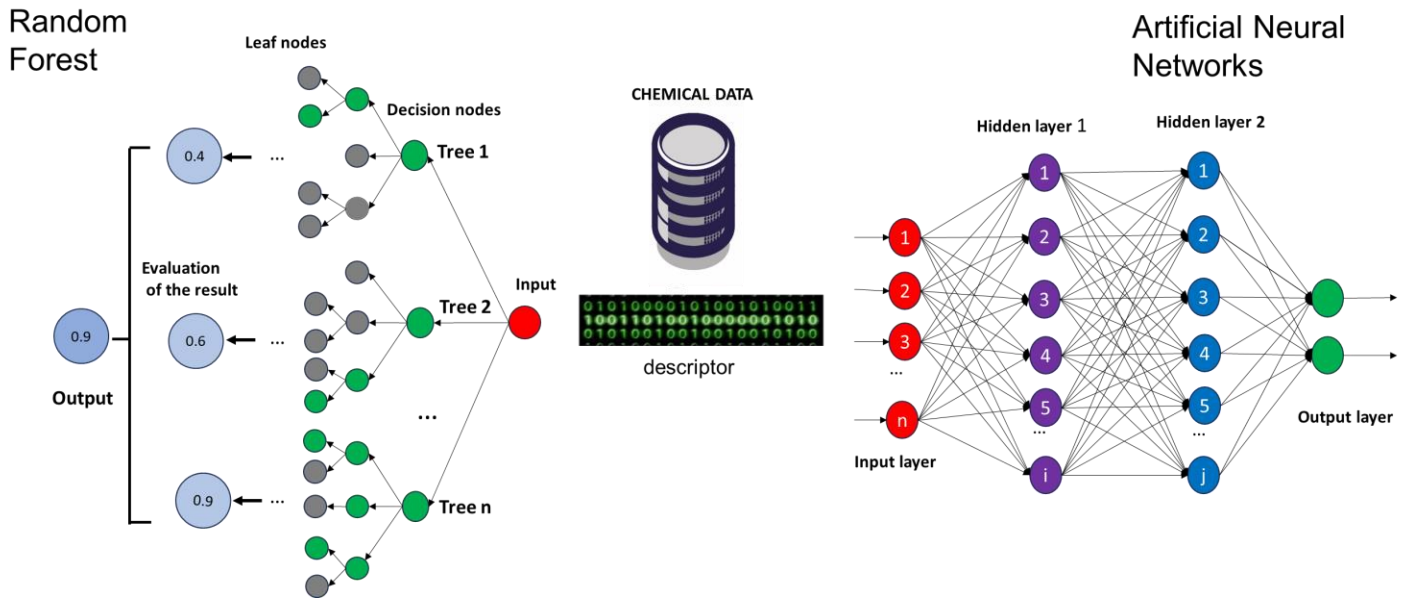

Figure 2. ML learning algorithms that are commonly used in the field of drug design: Random Forest (RF) and Artificial Neural Networks (ANNs).

RF is a supervised machine learning method built from a set (assembly) of decision tree (DT) algorithms, and they are organized based on a given priority of attributes or characteristics. They are used to solve regression and classification problems. They consist of two basic components: decision nodes and leaf nodes: Decision nodes represent the attributes or conjunctions of features that are used to analyze the data; and the leaf nodes represent the labels of each class.

On the other hand, ANNs are inspired by the biological neural networks of the human brain; that is, they are a cognitive computational system that is based on an algorithm represented by a matrix of interconnected nodes, so-called "neurons".[15] Each neuron incorporates several inputs and gives rise to a single output by means of a nonlinear function, affording the product of the inputs and parameters called "weights".[30,31] Since neurons are arranged in layers, the output of one layer becomes the input of the next. The weights of each neuron are adjusted during the training process in order to minimize a function that calculates a differential score between the overall output and the ideal output (Figure 2).[32,33] An example is the analysis of data points with specific characteristics, such as solubility, lipophilicity or bioavailability, to distinguish between high or low bioavailability, a relevant parameter in pharmacokinetics.

SVMs are a supervised methodology that facilitates the prediction of property values based on data classification and regression.[34] The goal of the SVM algorithm is to find a hyperplane that most properly separates different classes of data points. The classification task is accomplished by finding the differentiation criterion hyperplane between sets of points that belong to two distinct categories. First, the data are assigned to a new high-dimensional space in order to simplify the classification task. The decision limit is calculated using a protocol called "margin maximization", in which the distance between the hyperplane and the adjacent data points of each class is maximized. In the case of regression, the hyperplane is defined by optimizing the sum of the distances from the data points to the decision limit[34,35](Figure 3). SVMs are often used for binary properties

or activity potentials; for example, to discriminate between active and inefficient drugs, solubility in aqueous media, of synthetic feasibility, or to differentiate the specific activity of the compound.[36,37]

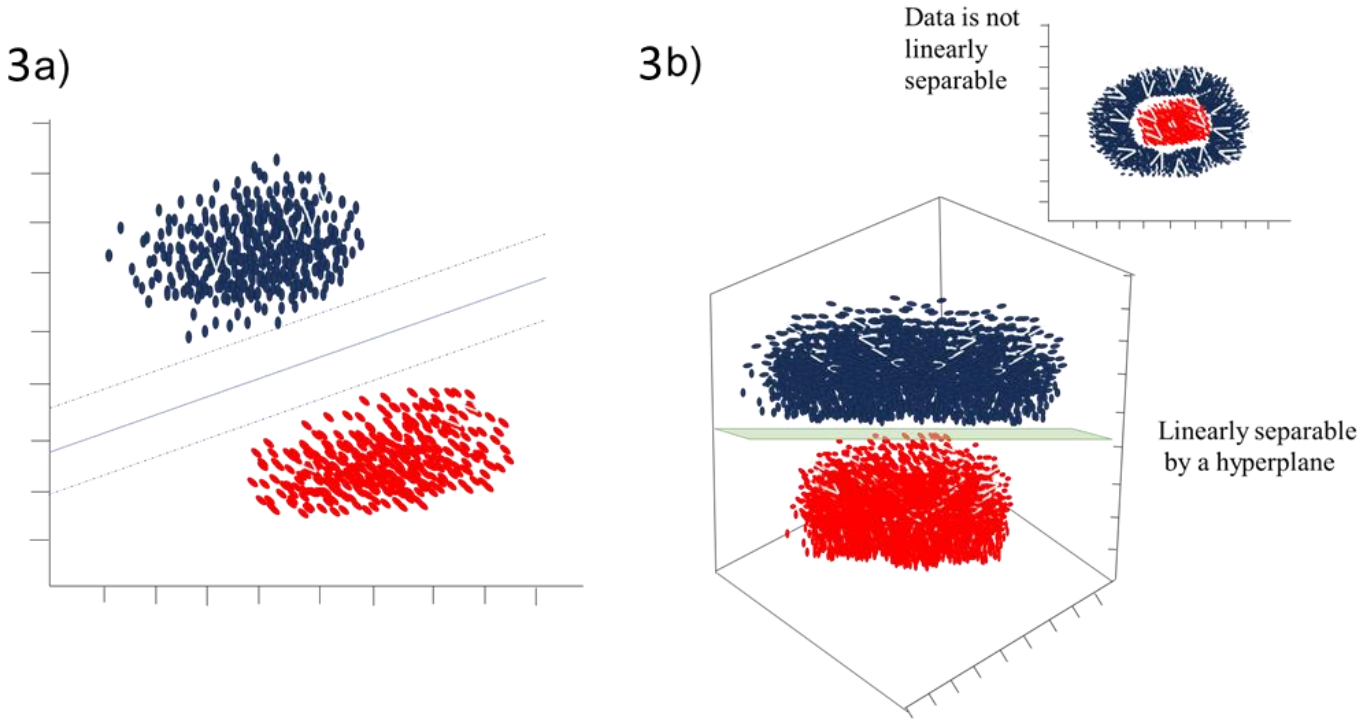

Figure 3. SVM classification examples. (a) Linearly separable datasets. The solid line indicates the hyperplane identified with the SVM technique that maximizes the distance to the nearest points of each class. Dashed lines delimit the margin for each class. b) Example: In the representation on top the datasets are not linearly separable, and in the representation at the bottom, the same datasets are shown, but they are mapped into a three-dimensional space. The data is now linearly separable by a hyperplane.[38]

The application of ML algorithms to a chemical target is a process that requires (1) the proper definition of the problem and the setting of specific objectives before making any structural decision. (2) The preparation of encoded input/output data, from a chemical descriptor or fingerprint; that is, the translation of chemical models into machine-readable representations.[14] (3) The selection of the ML architecture to which the study belongs (see below) as well as the selection of the algorithms necessary for its execution (RF, ANN, SVMs, etc.). Once the initial structure and associated datasets have been established, one can proceed with training, evaluation of the model, and finally the interpretation of the results.

Most ML techniques used in drug discovery are based on the construction of an output function from a data distribution or a probability density function p(*z*) generated from the structural information of the samples and the order of structural similarity (Figure 4).[14,39] The participation of these tasks is mainly presented in the virtual screening and are classified depending on the domain and can be generative or discriminative. A Generative Domain seeks to determine a joint probability function *G*(*x,y*)*, i.e.,* the probability of observing both the molecular representation *x*' (a representation analogous to that of the input data *x)* and its property (*y)* (e.g., solvation energy, enzyme affinity, etc.). A Discriminator Domain DD, on the other hand, aims to determine a conditional probability function *D*(*y/x*); that is, the probability of observing the property (*y*), given a molecular representation (*x*).[17,40]

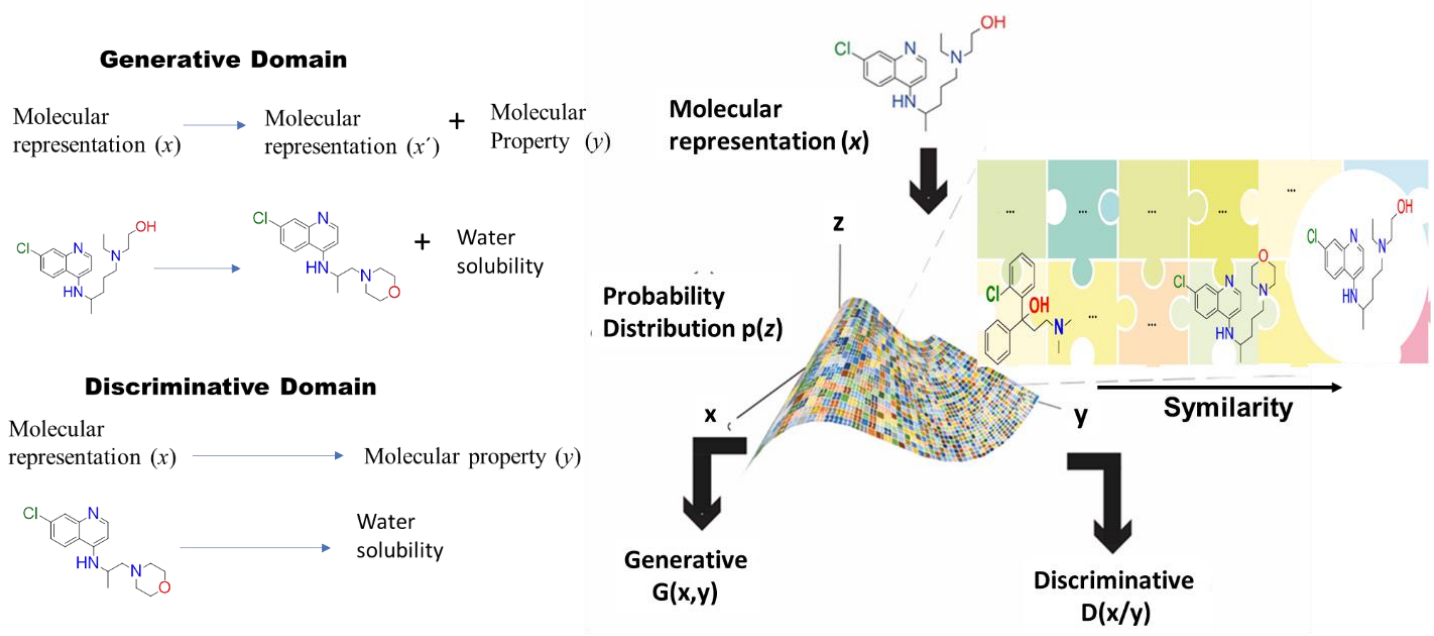

Figure 4. The Generative Study Domain determines a joint probability distribution $G(x, y)$. Discriminative, it determines conditional probabilities $D(y|x)$ and is based similarity Search (the assumption that structurally similar molecules exhibit similar biological activity compared to different or less similar molecules).[41]

3. "Smart" Virtual Screening

Virtual screening (VS) refers to a set of computational methods that helps to address specific problems by identifying potential successes through *in silico* experimentation. Today, this AI-supported technique offers considerable value in focusing the search, increasing hit rates through intelligent compound selection, and reducing time and costs to reach a satisfactory lead. It helps to select molecules within the virtual chemical space through mapping that allows the distribution of molecules and their properties to be visualized. The idea is to collect structural information from molecules within the explored space to search for bioactive compounds. Typically, this approach can be based on ligands (LBVS), on structural information (SBVS), and currently a third discipline is included, chemogenomics, an area that tries to unite both study spaces between ligands and drug targets (Figure 5).

LBVS seeks to estimate binding forces based on available ligand information (e.g., residue orientation, residue charges, etc.), anticipating that similar molecular structures can bind to a target protein and give rise to some activity. Within LBVS there are methodologies such as Similarity Search,[42] pharmacophore mapping,[43] or QSAR/QSRP.[44] SBVS collects information from a given three-dimensional molecular structure, as it assumes that the bonding of two molecules depends on the orientation of their atoms. SBVS models require a target structure and precise knowledge of the active sites that bind the ligands together. Numerous ML approaches can be used for the identification of binding sites, such as molecular dynamics (MD),[45] molecular coupling (MC)[46] and homology modeling (HM). [47-51] *Chemogenomics* works from an interdisciplinary approach that combines traditional ligand-based approaches with biological information on drug targets. The goal of chemogenomics is to understand the molecular recognition event between different ligands and potential drug targets. The binding protein and ligand shape have been previously studied as separate entities, whereas *chemigenomics* deals with datasets representing the relevant regions of the protein-ligand shared space.[52]

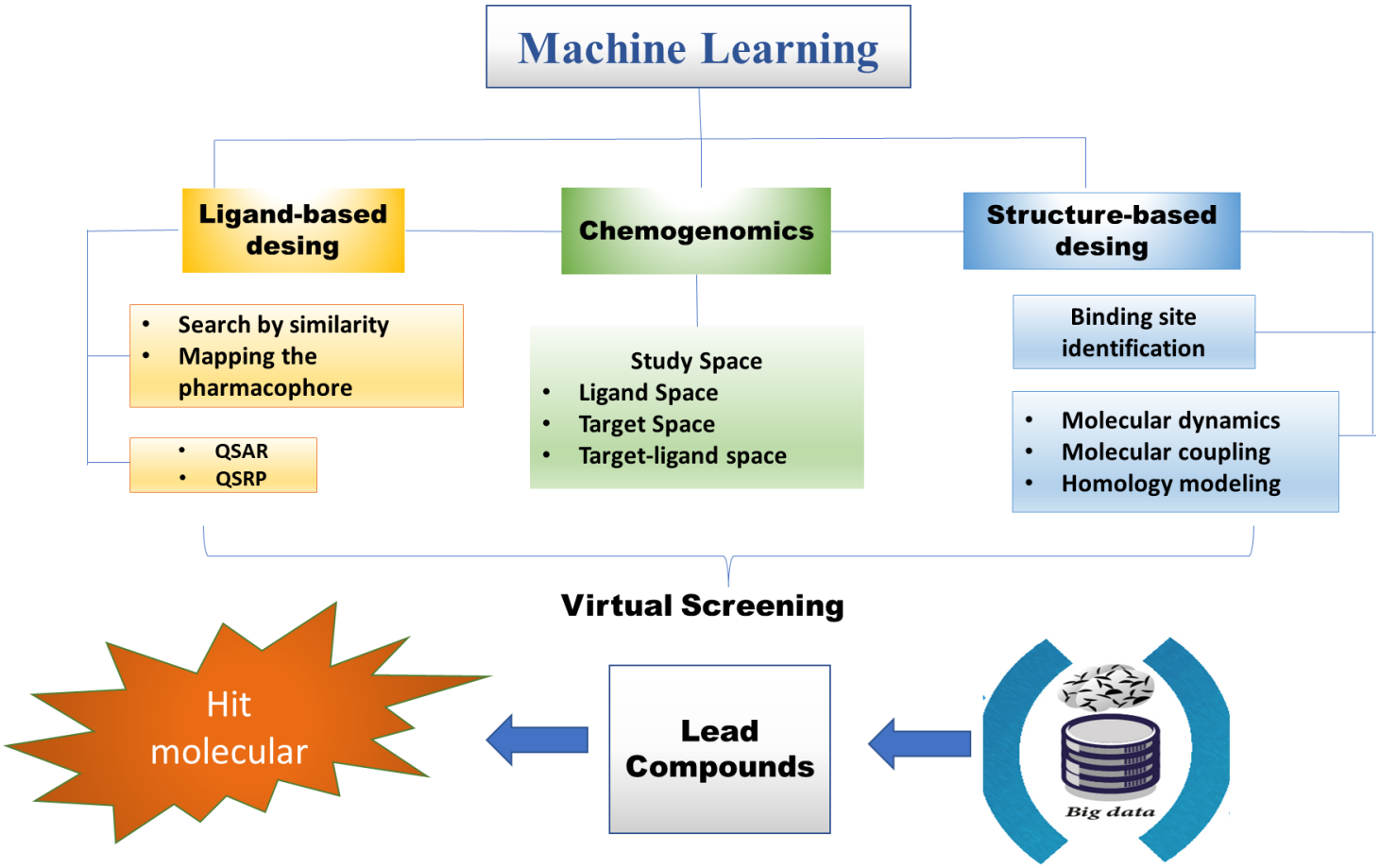

Figura 5. Intelligent Virtual Screening works by probing databases to obtain promising molecules "Molecular Hit".

3.1. Study space

3.1.1. Description of ligand spaces and drug targets. The ligand and protein space have been previously studied as separate entities, but chemogenomics studies deal with large datasets covering parts of the protein-ligand joint space. Since chemogenomics is concerned with the macromolecules with which ligands interact, it is of interest to develop ML strategies to visualize protein-ligand subspaces.[52] Basic assumptions of chemogenomics-based strategies are: (i) substrates that present some chemical similarity are likely to share the same targets and (ii) targets that accept similar ligands are likely to present similar structural patterns at the corresponding binding sites.[42] Therefore, completing the theoretical chemogenomics matrix implies that data on the "unbound" targets should be collected from the nearest neighboring "linked" targets, and that data on the "untargeted" ligands should be obtained from the nearest "targeted" ligands.

3.1.2. Ligand space. To navigate efficiently in ligand space it is necessary to encode a molecule into a descriptor (numerical values that characterize molecular properties),[53] and then use a metric that describes the similarity between different descriptors, which allows the classification of compounds in a virtual library according to their similarity to experimentally verified assets (an example is the Tanimoto coefficient,[54] which is known to be one of the best performers). Descriptors are typically classified according to their dimensionality, ranging from one-dimensional (1-D) to four-dimensional (4-D) properties.[55,56]

1-D descriptors are readily calculated, describe fundamental properties, such as molecular weight, number of atoms, and bonds, which can be derived from chemical formulas, and which can be used to predict pharmacokinetic properties, such as aqueous solubility,[57] the partition coefficient of 1-octanol-water,[58] plasma protein binding or bioavailability,[59] and also to classify compounds as drugs *vs*. non-drugs[60-62] or to design multiple ligands.[63]

Descriptors for relevant ligands are included in the 2-D family of topological descriptors, where connectivity (list of atoms and bonds) is presented and analyzed to encode salient atomic and bonding characteristics. Among these are methods based on 2D fingerprints (binary chain with a list of substructures or other predefined patterns).[64] Its essence lies in the search for specific properties in a molecule (molecular properties such as: donor potential hydrogen bonding, potential hydrogen bond acceptor, volume, and electropositivity[65]). The presence or absence of these properties is encoded in the form of an atomic key which is a four-bit string (sequence of digits "0" and "1").[66]

3-D fingerprint methods are more appropriate for similarity searches as they encode specific properties of the structural conformation (atomic coordinates, 3D pharmacophores, shapes, potentials, fields, spectra; etc.). Likewise, the description of molecular structure by 4-D models is considered for a set of conformers (conditionally, the fourth dimension), rather than a fixed conformation, as well as for describing its size and chemical structure.[67] Other methods used in chemoinformatics are SMILES (Simplified Molecular Line Input Specification),[68] WLN (Wiswesser Line Notation) or SDF (Structure Data Format) file,[60] which are representations of 2-D or 3-D chemical structures.

3.1.3. Target space. Proteins are commonly classified according to their sequence and structure. Numerous chemogenomics approaches apply the classification of target families (such as ion channels, kinases, G-protein-coupled receptors "GPCRs") or protein subfamilies (such as purinergic GPCRs) without taking into account the similarities of putative ligand binding sites.[70] To understand the structural architecture of the target, it is usually helpful to analyze the 2-D structure (presence of α-helices, δ-sheets, coils and random structures[71]) and it is actually better to analyze the 3-D structure (atomic coordinates provided by X-ray diffraction, NMR or molecular modeling[72]) and/or the corresponding conformation.

Targets can also be classified according to their pharmacological behavior, in particular binding affinity for a set of ligands.[73]

3.1.4.Target-ligand space. For ML exploration of target-ligand interactions, an efficient method is the use of ANNs, due to its ability to recognize precise patterns between independent variables and dependent variables. ANNs are usually trained on hundreds of thousands of existing chemical structures to provide three interconnected functions: an encoder, a decoder, and a predictor. The process occurs in two stages: namely, a generative stage and a predictive stage. In a typical study, the generative stage takes place when the encoder maps the digital representation of the molecule (fingerprint or chemical descriptor) as a continuous vector of real values, known as a latent space or probability density function. On the other hand, the predictive stage is designed to estimate the properties of molecules from the continuous vector representation of the molecule based on the assumption that compounds with similar structural patterns must have similar "drug-likeness" properties.[74] With the associated data, it is possible to navigate directly into the protein-ligand space, through complete arrays in which structural or affinity information is stored. Typically in an *xy* coordinate system, experimental evaluation of *y*-axis targets are variables that depend on compounds on the *x*-axis (e.g., in an *in vitro* affinity assay) affords an *xy* number array (IC50 values), which are most useful to predict the affinity of a new compound with an existing target using linear regression,[75] or, to measure a QSAR distance between two compounds or to estimate ADME properties and side effects.[76] Of particular interest are the Structural Interaction Fingerprints (IFPs),[77,78] that convert the atomic coordinates of a protein-ligand complex into a string

of bits that presents, for each residue of a binding site, the type of molecular interactions (e.g., hydrogen-bond, aromatic interaction, hydrophobic interaction) originated by a cocrystallized or coupled ligand.[73]

Representative models within the target-ligand space are the QSAR-based computational model,[79] which is based on mathematical models that make it possible to establish a relationship between the molecular structure and its chemical properties through various descriptors. That is, it seeks to find the correlation between molecular descriptors, fingerprints, graphs or other mathematical representations of molecules with real properties such as biological activity, ADMET, binding free energies, and kinetic data for the formation of protein-ligand complexes.[80,81] In this context, the chemical-biological properties are fundamentally determined by the molecular structure, which emphasizes the importance of studying their molecular properties. To correlate some relevant properties of molecules, such as melting point, boiling point, solubility, λ max, etc. where QSPR models are usually employed.[82,83] The quantitative structure-toxicity relationship (QSTR) model is relevant for the rejection of toxic molecules based on their parameters (TD50 and LD50 toxic and lethal dose parameters) in the early phases of drug development, thus increasing the quality of the selected candidates.[84] An illustrative example is the work carried out by Sepp Hochreiter et al., which consisted of the prediction of toxicity using the DeepTox protocol adjusted to ANNs, based on patterns of substructures previously reported as toxicophores (A toxicophore is a molecular moiety that is related to the toxic properties of a chemical substance).[85] The molecular biodegradability is also frequently associated with its molecular structure, and therefore, QSBR models are valuable for studying the biodegradability of a molecule in the context of environmental protection (Figure 6).[86]

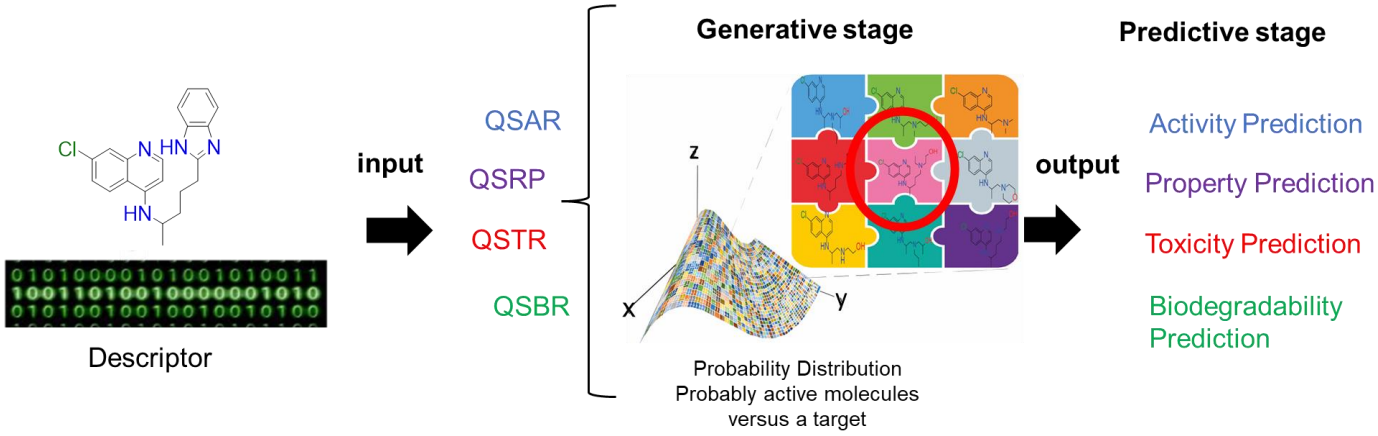

Figure 6. Artificial Neural Networks consist of a generative and a predictive stage, which aim to predict molecular properties by analysis by QSRP/QSAR/QSRT/QSBR by structural pattern recognition.

3.2. Molecular chirality in virtual screening

For VS by ML, it is important to consider the three-dimensional environment of the molecules. Chirality is the stereochemical property that describes the spatial arrangement of chemical substituents around a tetrahedral atom. Thus, a carbon atom is a center of chirality when it presents four non-similar substituents oriented to the vertices of the tetrahedron resulting from its sp^3 hybridization. Enantiomers (two molecules with the same chemical composition and connectivity, but different arrangement of the atoms around them) are non-superimposable mirror images.[87] These structures exhibit similar chemical properties (e.g., boiling/melting points), electron energies, and solubility, but they behave differently when interacting with external chiral environments.[88] Therefore, chirality is crucial for drug design, as protein-ligand interactions are highly influenced by ligand chirality, because proteins are also chiral.[89]

Most neural networks are designed to specify the chirality of potential drugs.[88] For example, Tyler Derr et al. developed a model called the Molecular-Kernel Graph Neural Network (MolKGNN), for ML training of three-dimensional molecular representation, and it is based on the design of a convolution of molecular graphs that captures patterns in structural representations (Figure 7). In particular, the structural nuclei (atoms) and their neighbors in the recipient molecule are compared with atoms learned by the algorithm (supervised learning). On the one hand, the data must be labeled with three-dimensional similarity attributes (bond distances, bond order, *R* or *S stereocenters*) and are evaluated with specific criteria, assigning them a score that depends on the degree of similarity.[90] As it becomes a multilayer system as the algorithm is trained, the pattern recognition power will be more efficient to the point that the three-dimensional characteristics of the molecule can be detected with certainty.

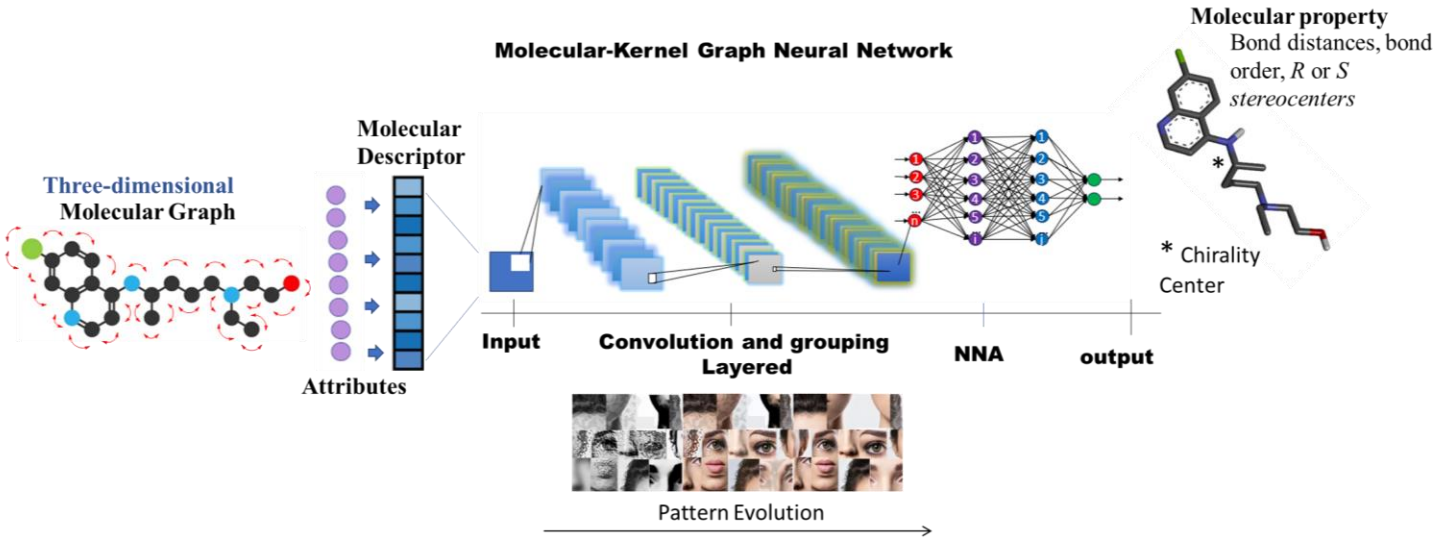

Figure 7. Artificial Neural Networks based on graph convolution. Molecular graph convolution takes patterns from 3D or 4D descriptors as input data labeled with similarity attributes (bond distances, bond order, *R* or *S* stereocenters) and as the algorithm learns the 3D patterns which enhance structural recognition.

4. Drug Design

After decades of intense research in academia, organic synthesis is recognized as an art that requires creativity and rigorous training.[91] AI drug development methods are very attractive because of the benefits they bring, such as saving resources, increasing chemical yields, as well as their potential to suggest more effective synthetic compounds.[92]

4.1. Computer-aided synthesis planning

New synthetic computer-aided synthesis planning (CASP) techniques are changing the way people work in this field and aim to minimize manipulations and maximize convergence for maximum reliability and efficiency in the experimental realization of estimated synthetic pathways.[93-96]

A typical CASP system consists of four modules: (a) The Reaction Template Database, which stores known reactions (is developed by manual entry and automatic extraction from commercial and open access databases), so that it is more feasible to design an optimal retrosynthesis path. (b) The retrosynthesis module, which is a

program that compares the structure of an input molecule with known reactions within the template database and returns the best match. (c) The ML-based guidance module (e.g., RF or ANNs to predict synthetic pathways through discrimination and correlation) which is used to evaluate potential precursors as well as the feasibility of various synthetic pathways. Finally, (d) access to the database of commercially available compounds (Figure 8).[93,94,96]

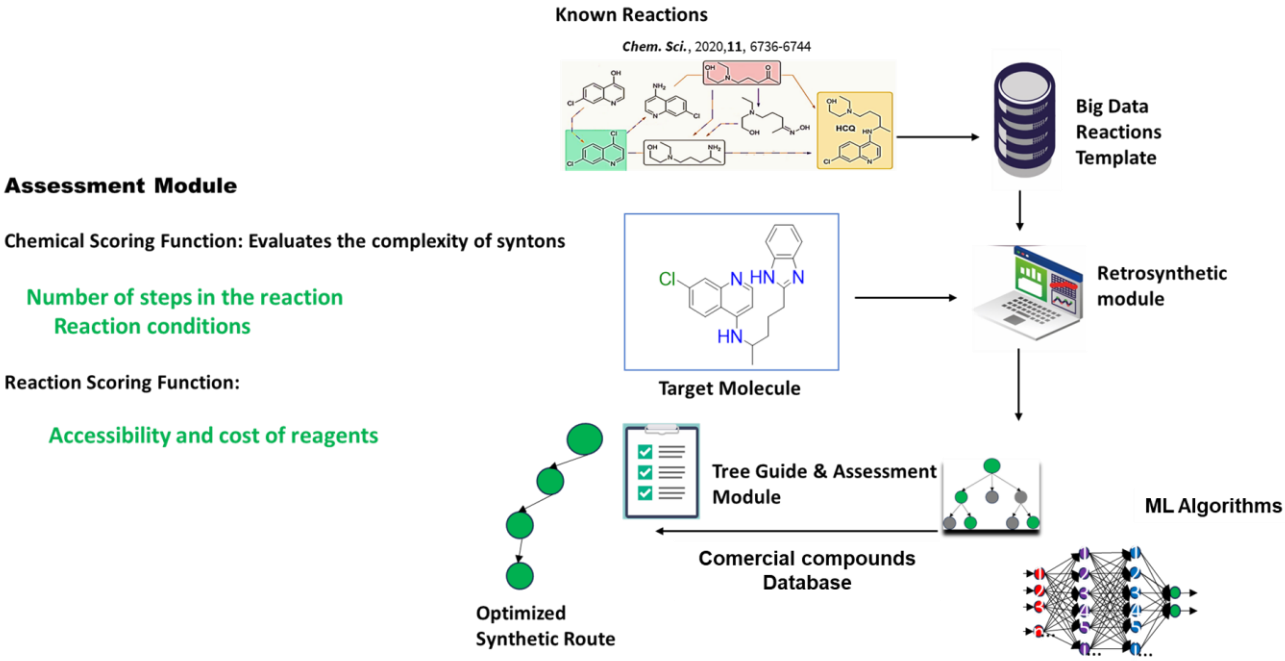

Figure 8. Main components of Computer-Aided Synthesis Planning (CASP).

Rule-based methods are conceptually similar to the procedures by which an organic chemist selects a known synthetic procedure to apply in the preparation of a specific synthetic target. These are generated from reaction templates contained in the database[95] and are evaluated with scoring functions: (a) the Chemical Scoring Function, to evaluate the difficulty of obtaining the desired molecules and (b) the reaction scoring function, which evaluates the cost of carrying out the proposed reactions. At this stage, synthetic routes that use fewer steps are favored and any conflict of reactivity or selectivity is penalized.[97,98] Structure-based and reaction-based scoring systems are currently being used here. Structure-based approaches assess the viability of molecular structure, examples include SYBA[99] that is based on a naïve Bayes classifier that is responsible for scoring individual fragments; RetroGNNS[100] ANN-based predicting molecules with biological activity and synthetic accessibility. On the other hand, reaction-based approaches predict synthetic accessibility by capturing the similarity of synthetic pathways stored in chemical reaction databases, e.g., dataset design,[101] SCScore[102] ANN-based.

An interesting case is the project developed by Grzybowski and his collaborators,[103] the software called SYNTHIA, which brings together modern high-power computing, artificial intelligence, and expert chemists' knowledge to design synthetic pathways that lead to specific targets, either previously synthesized or prepared for the first time. This program contains about 70,000 manually selected reaction transformation rules, which took the researchers more than 15 years to collect. However, it has been argued that SYNTHIA is impractical to manually code all synthetic pathways considering the exponential growth in the number of published reactions.[104] In addition, a simple template is usually not sufficient to reliably predict potentially useful reactions

because it only identifies reaction centers and their neighboring atoms without considering the overall information of the target molecule. An alternative method that tries to solve this problem is the so-called Molecular Transformer (autoregressive encoder-decoder model)[105,106] that provides higher accuracy and involves a lower computational cost. On the other hand, Zheng and collaborators[107] developed SCCROP, which suggests processes by retrosynthesis with an ANN-based syntax autocorrect, achieving 59% accuracy in a standard reference dataset outperforming rule-based methods. On the other hand, state-of-the-art methods such as LocalRetro divide generic reaction templates according to atomic change, bond polarity, and trains three different classifiers to obtain promising results for chemical synthesis.[108]

4.2. Molecular chirality in drug design

To examine regioselectivity and stereoselectivity, retrosynthesis predictors have been developed using SMILES and Molecular Transformer. However, SMILES may lead to erroneous predictions due to its fragile grammar. For example, a single character misplacement can be sufficient to invalidate an entire SMILES string, so any error within the molecular representation can lead to failure. To avoid these problems, alternative descriptors such as SELFIES[109] and DeepSMILES[110] have been used. A breakthrough within the molecular transformer that avoids inherent problems with the molecular descriptor was developed by Juyong Lee's group,[111] which replaces SMILES with atomic environments (AE),[112] making predictions with 58.3% accuracy.

The more appropriate incorporation of stereochemistry into molecular representations is a future direction being explored, as to date there are no reports of rule-based methods for the prediction of stereoselective reactions.[113] Graph-based methods systematically avoid stereochemistry issues, which highlights a great challenge.[114] In this context, the work carried out by Reymond's group,[115] based on Molecular Transformers, has the ability to interpret and predict reactions with stereochemical information of carbohydrates, where regio and stereoselectivity played an important role. Overall, they observed a consistent prediction accuracy above 70%, and validation was addressed by means of the experimental synthesis.

Despite CASP advances, the accuracy of predictions remains limited due to the overfitting of models to specific properties of training datasets, which can be mostly unbalanced or biased.[116] In the future, it is intended that CASP will be able to solve more challenging problems, such as chemo, regio, and stereoselective reactions. In this sense, CASP algorithms continue to evolve, particularly in the processing of template data that are based on reaction mechanisms and transition states.[96]

5. Applications of ML in the Laboratory

Many chemists and chemical engineers dream of the availability of a machine that has the ability to synthesize the molecules of interest, with no human participation.[117] Although recent advances in laboratory automation have reduced the time and effort required to perform manual chemical operations, the development of synthetic routes for the preparation of new molecules is still a manual process, requiring a large investment of time. However, it is hoped that the latest technological innovations in automation, robotics, and computing, as well as current advances in chemistry, synthesis and characterization of materials will be a catalyst to enable the autonomous development of chemical synthesis, both in industry and in academia.[118]

Since chemical laboratory research has always accumulated a large amount of experimental data on chemical and physical properties, reactions, chemical structures, and biological activities, autonomous labs promise to dramatically speed up the discovery process by improving automated experimentation platforms and decreasing measurement errors. However, in order to achieve this goal, the "levels of autonomy" must be taken

into account.[119] which are parameters that describe the degree of independence of human intervention in the autonomous laboratory. In addition, the application of standardized procedures with robotic support is intended to improve the reproducibility of experiments, reduce costs, synthesis times and analysis tests.[120,121]

5.1. Self-driving labs

It has been shown that self-driving labs (SDLs) can be based on simple natural language processing (NLP).[122] An SDL is a modular experimental platform assisted by machine learning that iteratively operates a series of experiments selected by the ML algorithm to achieve a user-defined goal.[123] SDL are built from a multidisciplinary establishment that combines AI with automated robotic platforms for the autonomous preparation of new molecules or materials. On the one hand, ML and modeling methods are used to predict the properties of materials being synthesized and suggest new experiments to improve the synthetic procedure. Robotics, computer vision, and automated characterization methods are used to perform the experiments and analyze the results. There are several salient examples of SDL. Christensen's group[124] developed an automated closed-loop system to carry out parallel experiments optimizing a Suzuki-Miyaura stereoselective coupling reaction, while other research groups developed continuous-flow chemical synthesis systems in C-C and C-N cross-coupling reactions, olefination, reductive amination, nucleophilic aromatic substitution (S_NAr), and photoredox catalysis,[125] or in lithium-halogen exchange reactions.[126] On the other hand, Granda's[127] group developed a reaction system controlled by an ML algorithm with the ability to explore the space of chemical reactions, an organic synthesis robot that can perform chemical reactions, and automated analysis (using nuclear magnetic resonance and infrared spectroscopy) as well as prediction of the reactivity of possible combinations of reactants. For last-generation laboratories habilitated to carry out high-pressure and high-temperature, Fast-Cat autonomously carries out Pareto-front mapping of homogeneous catalysts in gas-liquid reactions,[128] and Smart Dope, Autonomous Fluid Laboratory, for rapid synthesis and autonomous optimization of doping involving metal cations.[129] ChemOS, which is a flexible and modular system that manages the essential instructions for operating autonomous laboratories, from managing data collection, to the design of experimental procedures, and for providing instructions to robotic equipment. It also allows for the remote control of equipment, so that ChemOS can operate in different labs, including those located in different institutions.[130]

Other forms of virtual interaction have been reported in the form of a graphical user interface (GUI), which allows users to perform optimizations, monitor progress, and analyze results. Subsequent users of an optimized procedure only need to download an electronic file, comparable to a smartphone application, to implement the protocol on their own device.[131]

In this context, Coley et al.[132] in 2019 designed a synthesis planning module, based on ASKCOS, which is software that uses millions of reactions pulled from the United States Patent Office (USPTO), Reaxys, and other databases. It consists of a binary classifier and is intended to answer the question: Is there any set of conditions under which these reactants will give rise to the product of interest? Reactions that pass this filter with a user-adjustable threshold are added to an RF model. The optimal reaction conditions are provided by an ANN model trained to propose a prioritized list of most suitable reactants, solvents, catalysts, and temperature for such a transformation. The process consists of submodules implemented in a robotic flow chemistry platform (Figure 9).[133]

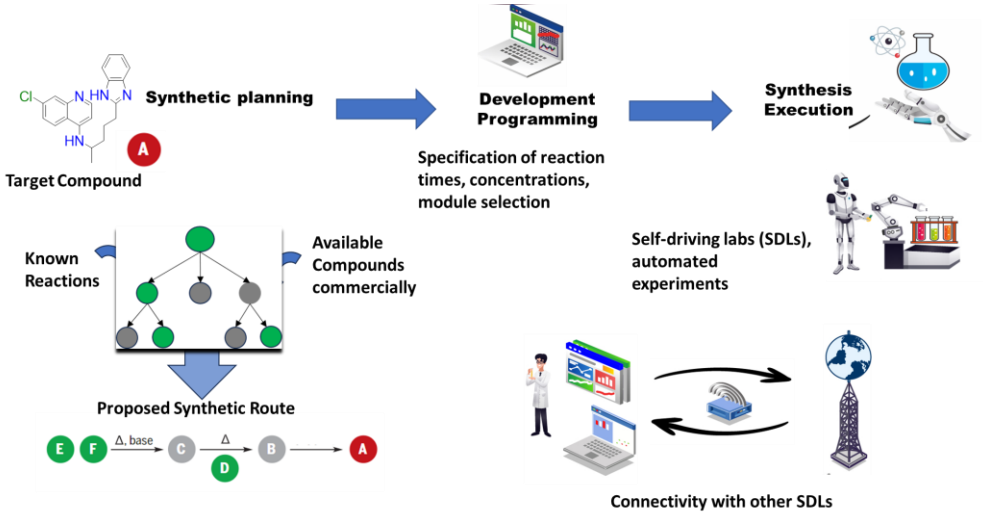

Figure 9. Synthesis planning module based on self-driving labs (SDLs); a robotic flow chemistry platform implemented in the synthesis of organic molecules.

Conclusions

Artificial intelligence and machine learning are becoming revolutionary technologies for the discovery and development of new drugs. This is due in part to efficient access to massive amounts of data, as well as advances in computer science, chemistry, medicine, engineering, and other fields. Scientific research, which is reflected in the increase in the number of articles published on this topic, and interest in industry will surely be reflected soon in the solution to crucial problems for humanity. Automation, on the other hand, will allow researchers to efficiently tackle a wider range of challenging problems, focusing on conceptualizing the results of the many experiments conducted and classified by intelligent computers, i.e., in other words, researchers cease to be responsible for the final design of the synthesis of interest.

References

1. Bender, A.; Cortes-Ciriano, I. *Drug Discovery Today* **2021**, *26*, 1040. https://doi.org/10.1016/j.drudis.2020.11.037
2. Won, J. H.; Lee, H. *Int. J. Mol. Sci.* **2021**, *22*, 5457. https://doi.org/10.3390/ijms22115457
3. Xu, Y.; Liu, X.; Cao, X.; Huang, C.; Liu, E.; Qian, S.; Liu, X.; Wu, Y.; Dong, F.; Qiu, C.-W.; Qiu, J.; Hua, K.; Su, W.; Wu, J.; Xu, H.; Han, Y.; Fu, C.; Yin, Z.; Liu, M.; Roepman, R.; Dietmann, S.; Virta, M.; Kengara, F.; Zhang, Z.; Zhang, L.; Zhao, T.; Dai, J.; Yang, J.; Lan, L.; Luo, M.; Liu, Z.; An, T.; Zhang, B.; He, X.; Cong, S.; Liu, X.; Zhang, W.; Lewis, J. P.; Tiedje, J. M.; Wang, Q.; An, Z.; Wang, F.; Zhang, L.; Huang, T.; Lu, C.; Cai, Z.; Wang, F.; Zhang, J. *Innovation* **2021**, *2*, 100179. https://doi.org/10.1016/j.xinn.2021.100179
4. Hughes, J. P.; Rees, S. S.; Kalindjian, S. B.; Philpott, K. L. *Br. J. Pharmacol.* **2011**, *162*, 1239. https://doi.org/10.1111/j.1476-5381.2010.01127.x
5. Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. *Expert Opin. Drug Discovery* **2021**, *16*, 949.

https://doi.org/10.1080/17460441.2021.1909567
6. Mao, J.; Akhtar, J.; Zhang, X.; Sun, L.; Guan, S.; Li, X.; Chen, G.; Liu, J.; Jeon, H. N.; Kim, M. S.; No, K. T.; Wang, G. *Science* **2021**, *24*, 103052. https://doi.org/10.1016/j.isci.2021.103052
7. Zhu, H. *Annu. Rev. Pharmacol. Toxicol.* **2020**, *60*, 573. https://doi.org/10.1146/annurev-pharmtox-010919-023324
8. Tormay, P. *Pharm. Med.* **2015**, *29*, 87. https://doi.org/10.1007/s40290-015-0090-x
9. Xue, H.; Li, J.; Xie, H.; Wang, Y. *Int. J. Biol. Sci.* **2018**, *14*, 1232. https://doi.org/10.7150/ijbs.24612
10. Martinez-Mayorga, K.; Rosas-Jiménez, J. G.; Gonzalez-Ponce, K.; López-López, E.; Neme, A.; Medina-Franco, J. L. *Chem. Sci.* **2024**, *15*, 1938. https://doi.org/10.1039/D3SC05534E
11. Miljković, F.; Medina-Franco, J. L. *Artif. Intell. Life Sci.* **2024**, *5*, 100096. https://doi.org/10.1016/j.ailsci.2024.100096
12. Xie, Y.; Sattari, K.; Zhang, C.; Lin, J. *Prog. Mater. Sci.* **2023**, *132*, 101043. https://doi.org/10.1016/j.pmatsci.2022.101043
13. Jason B. *Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data Transforms in Phyton*; Machine Learning Mastery, 2020. pp 2.
14. Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. *Drug Discovery Today* **2018**, *23*, 1538. https://doi.org/10.1016/j.drudis.2018.05.010
15. Su, M.; Liang, B.; Ma, S.; Alzubi, J.; Nayyar, A.; Kumar, A. *J. Phys.: Conf. Ser.* **2018**, *1142*, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012
16. Sarker, I. H. *SN Comput. Sci.* **2022**, *3*, 158. https://doi.org/10.1007/s42979-022-01043-x
17. Maglogiannis, I.; Karpouzis, K.; Wallace, M. *Emerging Artificial Intelligence Applications in Computer Engineering : Real Word AI Systems with Applications in eHealth*, Information Retrieval and Pervasive Technologies.; IOS Press: BG Amsterdam, 2007; Vol. 160, pp 281.
18. Sellwood, M. A.; Ahmed, M.; Segler, M. H. S.; Brown, N. *Future Med. Chem.* **2018**, *10*, 2025. https://doi.org/10.4155/fmc-2018-0212
19. Chu, X.; Lin, Y.; Wang, Y.; Wang, L.; Wang, J.; Gao, J. *IJCAI Int. Jt. Conf. Artif. Intell.* **2019**, *2019-August*, 4518. https://doi.org/10.24963/ijcai.2019/628
20. Zhou, Z.; Li, X.; Zare, R. N. *ACS Cent. Sci.* **2017**, *3*, 1337. https://doi.org/10.24963/ijcai.2019/628
21. Popova, M.; Isayev, O.; Tropsha, A. *Sci. Adv.* **2018**, *4*, eaap7885. https://doi.org/10.1126/sciadv.aap7885
22. Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. *J. Cheminf.* **2017**, *9*, 48. https://doi.org/10.1186/s13321-017-0235-x
23. Burzykowski, T.; Geubbelmans, M.; Rousseau, A.-J.; Valkenborg, D. *Am. J. Orthod. Dentofacial Orthop.* **2023**, *164*, 295. https://doi.org/10.1016/j.ajodo.2023.05.007
24. Mata, J.; Salazar, F.; Barateiro, J.; Antunes, A. *Water* **2021**, *13*, 19. https://doi.org/10.3390/w13192717

25. Inamuddin, Z.; Altalhi, T.; Cruz, J. N.; El-Deen Refat, M. S. *Drug Design using Machine Learning* **2022**, *1*. pp 39-51. https://doi.org/10.1002/9781394167258
26. Ongsulee, P. *International Conference on ICT and Knowledge Engineering* **2018**, *1*. https://doi.org/10.1109/ICTKE.2017.8259629
27. Cover, T. M.; Hart, P. E. *IEEE Trans. Inf. Theory* **1967**, *13*, 21. https://doi.org/10.1109/TIT.1967.1053964
28. Friedman, N.; Geiger, D.; Goldszmidt, M. *Machine Learning* **1997**, *29*, 131. https://doi.org/10.1023/A:1007465528199
29. Schultz, C.; Alegría, A. C.; Cornelis, J.; Sahli, H. *Applied Geography* **2016**, *66*, 52. https://doi.org/10.1016/j.apgeog.2015.11.005
30. Nandi, S.; Bagchi, M. C. *Chem. Biol. Drug Des.* **2011**, *78*, 587. https://doi.org/10.1111/j.1747-0285.2011.01177.x
31. Goudarzi, N.; Goodarzi, M.; Chen, T. *Med. Chem. Res.* **2012**, *21*, 437. https://doi.org/10.1007/s00044-010-9542-8
32. Gertrudes, J. C.; Maltarollo, V. G.; Silva, R. A.; Oliveira, P. R.; Honorio, K. M.; da Silva, A. B. F. *Curr. Med. Chem.* **2012**, *19*, 4289. https://doi.org/10.2174/092986712802884259
33. Milac, A. L.; Avram, S.; Petrescu, A. J. *J. Mol. Graphics Modell.* **2006**, *25*, 37. https://doi.org/10.1016/j.jmgm.2005.09.014
34. Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048. https://doi.org/10.1021/ci0340916
35. Hastie, T.; Tibshirani, R.; Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction; Springer New York, 2nd Edition, **2009**, pp 9-41.
36. Podolyan, Y.; Walters, M. A.; Karypis, G. *J. Chem. Inf. Model.* **2010**, *50*, 979. https://doi.org/10.1021/ci900301v
37. Jorissen, R. N.; Gilson, M. K. *J. Chem. Inf. Model.* **2005**, *45*, 549. https://doi.org/10.1021/ci049641u
38. Brereton, R. G.; Lloyd, G. *R. Analyst* **2010**, *135*, 230. https://doi.org/10.1039/B918972F
39. Yang, J.; Cai, Y.; Zhao, K.; Xie, H.; Chen, X. *Drug Discovery Today* **2022**, *27*, 103356. https://doi.org/10.1016/j.drudis.2022.103356
40. Torne, L.; Binns, R. *Drug Discovery Today* **2018**, *23*, 1922. https://doi.org/10.1016/j.drudis.2018.09.005
41. Kubinyi, H. *Perspect. Drug Discovery Des.* **1998**, *9*, 225. https://doi.org/10.1023/A:1027221424359
42. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R. K. *Drug Discovery Today* **2021**, *26*, 80. https://doi.org/10.1016/j.drudis.2020.10.010
43. Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. *BMC Bioinf.* **2014**, *15*, 291. https://doi.org/10.1186/1471-2105-15-291
44. Raschka, S.; Kaufman, B. *Methods* **2020**, *180*, 89. https://doi.org/10.1016/j.ymeth.2020.06.016
45. Wang, Y.; Lamim Ribeiro, J. M.; Tiwary, P. *Curr. Opin. Struct. Biol.* **2020**, *61*, 139.

https://doi.org/10.1016/j.sbi.2019.12.016
46. Khamis, M. A.; Gomaa, W.; Ahmed, W. F. *Artif. Intell. Med.* **2015**, *63*, 135. https://doi.org/10.1016/j.artmed.2015.02.002
47. Lavecchia, A. *Drug Discovery Today* **2015**, *20*, 318. https://doi.org/10.1016/j.drudis.2014.10.012
48. Yang, J.; Shen, C.; Huang, N. *Front. Pharmacol.* **2020**, *11*, 508760. https://doi.org/10.3389/fphar.2020.00069
49. Verma, N.; Qu, X.; Trozzi, F.; Elsaied, M.; Karki, N.; Tao, Y.; Zoltowski, B., Larson, E. C.; Kraka, E. *Int. J. Mol. Sci.* **2021**, *22*, 1392. https://doi.org/10.3390/ijms22031392
50. Zhao, J.; Cao, Y.; Zhang, L. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 417. https://doi.org/10.1016/j.csbj.2020.02.008
51. Whitfield, T. W.; Ragland, D. A.; Zeldovich, K. B.; Schiffer, C. A. *J. Chem. Theory Comput.* **2020**, *16*, 1284. https://doi.org/10.1021/acs.jctc.9b00781
52. Strömbergsson, H.; Kleywegt, G. J. *BMC Bioinf.* **2009**, *10*, S13. https://doi.org/10.1186/1471-2105-10-S6-S13
53. Andrew, R.; Gillet V. J. An Introduction to Chemoinformatics, Springer, 2003, pp 126
54. Rácz, A.; Bajusz, D.; Héberger, K. *J. Cheminf.* **2018**, *10*, 48. https://doi.org/10.1186/s13321-018-0302-y
55. Sawada, R.; Kotera, M.; Yamanishi, Y. *Mol. Inf.* **2014**, *33*, 719. https://doi.org/10.1002/minf.201400066
56. Kuz'min, V. E.; Artemenko, A. G.; Polischuk, P. G.; Muratov, E. N.; Hromov, A. I.; Liahovskiy, A. V.; Andronati, S. A.; Makan, S. Y. *J. Mol. Model.* **2005**, *11*, 457. https://doi.org/10.1007/s00894-005-0237-x
57. Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. *Mutagenesis* **2004**, *19*, 365. https://doi.org/10.1093/mutage/geh043
58. Ehresmann, B.; De Groot, M. J.; Clark, T. *J. Chem. Inf. Model.* **2005**, *45*, 1053. https://doi.org/10.1021/ci050025n
59. Wu, W.; Wang, Y.; Que, L. *Eur. J. Pharm. Biopharm.* **2006**, *63*, 288. https://doi.org/10.1016/j.ejpb.2005.12.005
60. Hu, Q.; Feng, M.; Lai, L.; Pei, *J. Front. Genet.* **2018**, *9*, 422486. https://doi.org/10.3389/fgene.2018.00585
61. Schierz, A. C.; King, R. D. Drugs and Drug-Like Compounds: Discriminating Approved Pharmaceuticals from Screening-Library Compounds. In Pattern Recognition in Bioinformatics: Kadirkamanathan, V.; Sanguinetti, G.; Girolami, M.; Niranjan, M.; Noirel, J. (eds). PRIB 2009. Lecture Notes in Computer Science, vol 5780. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04031-3_29
62. Sadowski, J.; Kubinyi, H. *J. Med. Chem.* **1998**, *41*, 3325. https://doi.org/10.1021/jm9706776
63. Morphy, R.; Rankovic, Z. *J. Med. Chem.* **2006**, *49*, 4961. https://doi.org/10.1021/jm0603015
64. Yang, H.; Li, J.; Wu, Z.; Li, W.; Liu, G.; Tang, Y. *Chem. Res. Toxicol.* **2017**, *30*, 1355. https://doi.org/10.1021/acs.chemrestox.7b00083

65. Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391. https://doi.org/10.1021/ci025569t
66. Cammarata, A.; Menon, G. K. *J. Med. Chem.* **1976**, *19*, 739. https://doi.org/10.1021/jm00228a001
67. Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526. https://doi.org/10.1021/ci049898s
68. Weininger, D. *J. Chem. Inf. Comput. Sci.***1988**, *28*, 31. https://doi.org/10.1021/ci00057a005
69. What is the correct format for compounds in SDF or MOL files? - Progenesis SDF Studio https://www.nonlinear.com/progenesis/sdf-studio/v0.9/faq/sdf-file-format-guidance.aspx (accessed Nov 29, 2023).
70. Klabunde, T. *Br. J. Pharmacol.* **2007**, *152*, 5. https://doi.org/10.1038/sj.bjp.0707308
71. Moore, S. J.; Deplazes, E.; Mancera, R. L. *Proteins: Struct., Funct., Bioinf.* **2023**, *91*, 338. https://doi.org/10.1002/prot.26432
72. Nachiappan, M.; Guru Raj Rao, R.; Richard, M.; Prabhu, D.; Rajamanikandan, S.; Chitra, J. P.; Jeyakanthan, J. Molecular Docking for Computer-Aided Drug Design: Fundamentals, Techniques, Resources and Applications; London, UK : Academic Press, 2021, pp 119-140. https://doi.org/10.1016/B978-0-12-822312-3.00007-2
73. Rognan, D. *Br. J. Pharmacol.* **2007**, *152*, 38. https://doi.org/10.1038/sj.bjp.0707307
74. Sanchez-Lengeling, B.; Aspuru-Guzik, A. *Science* **2018**, *361*, 6400. https://doi.org/10.1126/science.aat2663
75. Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. *Chem. Biol.* **1995**, *2*, 107. https://doi.org/10.1016/1074-5521(95)90283-X
76. Krejsa, C.; Horvath, D.; Rogalski, S. L.; Penzotti, J.; Mao, B.; Barbosa, F.; Migeon, J. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 4.
77. Szulc, N. A.; Mackiewicz, Z.; Bujnicki, J. M.; Stefaniak, F. *Briefings Bioinf.* **2023**, *24*, 4. https://doi.org/10.1093/bib/bbad187
78. Deng, Z.; Chuaqui, C.; Singh, J. *J. Med. Chem.* **2004**, *47*, 337. https://doi.org/10.1021/jm030331x
79. Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. *Chem. Sci.* **2021**, *12*, 7866. https://doi.org/10.1039/D0SC06805E
80. Soares, T. A.; Nunes-Alves, A.; Mazzolari, A.; Ruggiu, F.; Wei, G. W.; Merz, K. *J. Chem. Inf. Model.* **2022**, *62*, 5317. https://doi.org/10.1021/acs.jcim.2c01422
81. Cramer, R. D. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 35. https://doi.org/10.1007/s10822-011-9495-0
82. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69. https://doi.org/10.1002/qsar.200390007
83. Wadhwa, P.; Mittal, A. In Computer Aided Pharmaceutics and Drug Delivery; Springer, Singapore, 2022; pp 54.
84. Devillers, J. *SAR QSAR Environ. Res.* **2004**, *15*, 501.

https://doi.org/10.1080/10629360412331297443
85. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. *Front. Environ. Sci.* **2016**, *3*, 167215. https://doi.org/10.3389/fenvs.2015.00080
86. Lin, X.; Li, X.; Lin, X. *Molecules* **2020**, *25*, 6. https://doi.org/10.3390/molecules25061375
87. Juaristi, E. Introduction to Stereochemistry and Conformational Analysis, Wiley, New York, 1991.
88. Adams, K.; Pattanaik, L.; Coley, C. W. *arXiv* **2021**. https://doi.org/10.48550/arxiv.2110.04383
89. Nguyen, L. A.; He, H.; Pham-Huy, C. *Int. J. Biomed. Sci.* **2006**, *2*, 85.
90. Liu, Y.; Wang, Y.; Vu, O.; Moretti, R.; Bodenheimer, B.; Meiler, J.; Derr, T. *bioRxiv* **2022**, *8*, 24. https://doi.org/10.1101/2022.08.24.505155
91. Lippard, S. J. *Nature* **2002**, *416*, 587. https://doi.org/10.1038/416587a
92. Turzo, S. B. A.; Hantz, E. R.; Lindert, S. *QRB Discov.* **2022**, *3*, e14. https://doi.org/10.1017/qrd.2022.12
93. Wang, Z.; Zhao, W.; Hao, G.; Song, B. *Org. Chem. Front*. **2021**, *8*, 812. https://doi.org/10.1039/D0QO00946F
94. Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. *Nat. Catal.* **2021**, *4*, 98. https://doi.org/10.1038/s41929-020-00556-z
95. Lin, Y.; Zhang, R.; Wang, D.; Cernak, T. *Science* **2023**, *379*, 453. https://doi.org/10.1126/science.ade8459
96. Wang, Z.; Zhang, W.; Liu, B. *Chin. J. Chem.* **2021**, *39*, 3127. https://doi.org/10.1002/cjoc.202100273
97. Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. *A. Chem* **2018**, *4*, 522. https://doi.org/10.1016/j.chempr.2018.02.002
98. Wender, P. A.; Verma, V. A.; Paxton, T. J.; Pillow, T. H. *Acc. Chem. Res.* **2008**, *41*, 40. https://doi.org/10.1021/ar700155p
99. Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. *J. Cheminf.* **2020**, *12*, 35. https://doi.org/10.1186/s13321-020-00439-2
100. Liu, C. H.; Korablyov, M.; Jastrzębski, S.; Włodarczyk-Pruszyński, P.; Bengio, Y.; Segler, M. *J. Chem. Inf. Model.* **2022**, *62*, 2293. https://doi.org/10.1021/acs.jcim.1c01476
101. Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. *ACS Cent. Sci.* **2023**, *9*, 2196. https://doi.org/10.1021/acscentsci.3c01163
102. Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. *J. Chem. Inf. Model.* **2018**, *58*, 252. https://doi.org/10.1021/acs.jcim.7b00622
103. Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904. https://doi.org/10.1002/anie.201506101
104. Coley, C. W.; Green, W. H.; Jensen, K. F. *Acc. Chem. Res.* **2018**, *51*, 1281. https://doi.org/10.1021/acs.accounts.8b00087

105. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Bekas, C.; Lee, A. *ACS Cent. Sci.* **2019**, *5*, 1572. https://doi.org/10.1021/acscentsci.9b00576
106. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. *Advances in Neural Information Processing Systems* **2017**, 5999.
107. Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. *J. Chem. Inf. Model*. **2020**, *60*, 47. https://doi.org/10.1021/acs.jcim.9b00949
108. Chen, S.; Jung, Y. *JACS Au* **2021**, *1*, 1612. https://doi.org/10.1021/jacsau.1c00246
109. Krenn, M.; Häse, F.; Nigam, A. K.; Friederich, P.; Aspuru-Guzik, A. *Mach. Learn. Sci. Technol.* **2020**, *1*, 045024. https://doi.org/10.1088/2632-2153/aba947
110. O'Boyle, N. M.; Dalke, A. *ChemRxiv* **2018**. https://doi.org/10.26434/chemrxiv.7097960.v1
111. Ucak, U. V.; Ashyrmamatov, I.; Ko, J.; Lee, J. *Nat. Commun.* **2022**, *13*, 1186. https://doi.org/10.1038/s41467-022-28857-w
112. Hähnke, V. D.; Bolton, E. E.; Bryant, S. H. *J. Cheminf.* **2015**, *7*, 41. https://doi.org/10.1186/s13321-015-0076-4
113. Lin, K.; Xu, Y.; Pei, J.; Lai, L. *Chem. Sci.* **2020**, *11*, 3355. https://doi.org/10.1039/C9SC03666K
114. Zhang, C.; Lapkin, A. A. *React. Chem. Eng.* **2023**, *8*, 2491. https://doi.org/10.1039/D2RE00406B
115. Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J. L. *Nat. Commun.* **2020**, *11*, 4874. https://doi.org/10.1038/s41467-020-18671-7
116. Skoraczyński, G.; Kitlas, M.; Miasojedow, B.; Gambin, A. *J. Cheminf.* **2023**, *15*, 6. https://doi.org/10.1186/s13321-023-00678-z
117. Peplow, M. *Nature* **2014**, *512*, 20. https://doi.org/10.1038/512020a
118. Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C.; Brabec, C. J.; Maruyama, B.; Persson, K. A.; Aspuru-Guzik, A. *Nat. Rev. Mater.* **2018**, *3*, 5. https://doi.org/10.1038/s41578-018-0005-z
119. Martin, H. G.; Radivojevic, T.; Zucker, J.; Bouchard, K.; Sustarich, J.; Peisert, S.; Arnold, D.; Hillson, N.; Babnigg, G.; Marti, J. M.; Mungall, C. J.; Beckham, G. T.; Waldburger, L.; Carothers, J.; Sundaram, S. S.; Agarwal, D.; Simmons, B. A.; Backman, T.; Banerjee, D.; Tanjore, D.; Ramakrishnan, L.; Singh, A. *Curr. Opin. Biotechnol.* **2023**, *79*, 102881. https://doi.org/10.1016/j.copbio.2022.102881
120. Häse, F.; Roch, L. M.; Aspuru-Guzik, A. *Trends Chem.* **2019**, *1*, 282. https://doi.org/10.1016/j.trechm.2019.02.007
121. Fitzpatrick, D. E.; Battilocchio, C.; Ley, S. V. *ACS Cent. Sci.* **2016**, *2*, 131. https://doi.org/10.1021/acscentsci.6b00015
122. Caramelli, D.; Salley, D.; Henson, A.; Camarasa, G. A.; Sharabi, S.; Keenan, G.; Cronin, L. *Nat. Commun.* **2018**, *9*, 3406. https://doi.org/10.1038/s41467-018-05828-8
123. Abolhasani, M.; Kumacheva, E. *Nat. Synth.* **2023**, *2*, 483.

https://doi.org/10.1038/s44160-022-00231-0
124. Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. *Commun. Chem.* **2021**, *4*, 112. https://doi.org/10.1038/s42004-021-00550-x
125. Bédard, A. C.; Adamo, A.; Aroh, K. C.; Russell, M. G.; Bedermann, A. A.; Torosian, J.; Yue, B.; Jensen, K. F.; Jamison, T. F. *Science* **2018**, *361*, 1220. https://doi.org/10.1126/science.aat0650
126. Karan, D.; Chen, G.; Jose, N.; Bai, J.; McDaid, P.; Lapkin, A. A. *React. Chem. Eng.* **2024**, *9*, 619. https://doi.org/10.1039/D3RE00539A
127. Granda, J. M.; Donina, L.; Dragone, V.; Long, D. L.; Cronin, L. *Nature* **2018**, *559*, 377. https://doi.org/10.1038/s41586-018-0307-8
128. Bennett, J. A.; Orouji, N.; Khan, M.; Sadeghi, S.; Rodgers, J.; Abolhasani, M. *Nat. Chem. Eng.* **2024**, *1*, 240. https://doi.org/10.1038/s44286-024-00033-5
129. Bateni, F.; Sadeghi, S.; Orouji, N.; Bennett, J. A.; Punati, V. S.; Stark, C.; Wang, J.; Rosko, M. C.; Chen, O.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. *Adv. Energy Mater.* **2024**, *14*, 2302303. https://doi.org/10.1002/aenm.202302303
130. Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. *Sci. Robot.* **2018**, *20*, 3. https://doi.org/10.1126/scirobotics.aat5559
131. Bédard, A. C.; Adamo, A.; Aroh, K. C.; Russell, M. G.; Bedermann, A. A.; Torosian, J.; Yue, B.; Jensen, K. F.; Jamison, T. F. *Science* **2018**, *361*, 1220. https://doi.org/10.1126/science.aat0650
132. Collins, N.; Stout, D.; Lim, J. P.; Malerich, J. P.; White, J. D.; Madrid, P. B.; Latendresse, M.; Krieger, D.; Szeto, J.; Vu, V. A.; Rucker, K.; Deleo, M.; Gorfu, Y.; Krummenacker, M.; Hokama, L. A.; Karp, P.; Mallya, S. *Org. Process Res. Dev.* **2020**, *24*, 2064. https://doi.org/10.1021/acs.oprd.0c00143
133. Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; John Hart, A.; Jamison, T. F.; Jensen, K. F. *Science* **2019**, *365*, 6453. https://doi.org/10.1126/science.aax1566

Authors' Biographies

Carlos Naranjo earned his bachelor's degree in pharmaceutical and biological chemistry from the National Autonomous University of Mexico (2019). Presently he is doctoral student at CINVESTAV-IPN, Mexico, in the area of organic chemistry under the direction of Professor Eusebio Juaristi. His research focuses on asymmetric organocatalysis and mechanochemistry. Carlos Naranjo is particularly interested in artificial intelligence and medicinal chemistry.

Carlos A. Coello Coello received a PhD in Computer Science from Tulane University in 1996. His research has mainly focused on the design of new multi-objective optimization algorithms based on bio-inspired metaheuristics (e.g., evolutionary algorithms), which is an area in which he has made pioneering contributions. He has received several awards, including the 2012 Presidential Medal of Science in Physics, Mathematics and Natural Sciences from Mexico's presidency, the 2013 IEEE Kiyo Tomiyasu Award, the 2016 The World Academy of Sciences (TWAS) Award in "*Engineering Sciences*", and the 2021 IEEE Computational Intelligence Society Evolutionary Computation Pioneer Award. In May 2023, he became a member of "El Colegio Nacional", which is the highest academic distinction in Mexico.

Eusebio Juaristi studied chemistry at Tecnológico de Monterrey (B. Sc., 1972) and at the University of North Carolina at Chapel Hill (Ph.D., 1977). Juaristi became a postdoctoral associate at the University of California in Berkeley (1977-1978) and research associate at Syntex, Palo Alto, California (1978-1979) before returning to Mexico where he is now Professor of Chemistry at CINVESTAV-IPN. Juaristi was Visiting Professor at the E.T.H.-Zurich, 1985-1986 and 1992-1993, at the University of California in Berkeley (1999-2000), and at RWTH-Aachen, Germany (May-July 2013). *Scientific contributions*. Physical organic chemistry with emphasis in conformational analysis and stereochemistry, for example in the study of the *anomeric effect*. Juaristi has also worked in the areas of asymmetric synthesis, in particular on enantioselective synthesis of β-amino acids. Other chemistry areas where Juaristi has had influence include applications of computational chemistry, asymmetric organocatalysis, and sustainable ("green") chemistry. *Awards*. Medal of the Mexican Academy of Sciences for Young Scientists in 1988; National Chemistry Award granted by the Mexican Chemical Society in 1994; and the Presidential Medal in Sciences and Arts in 1998. In February of 2006 he became a Member of "El Colegio Nacional", highest academic distinction in Mexico. In the year 2012 he received the prestigious Georg Forster Award of the Alexander von Humboldt Foundation.