Machine Learning - driven insights for predicting the impact of nanoparticles on the functionality of biomolecules, Illustrated by the case of DNA Damage-Inducible Transcript 3 (CHOP) inhibitors

Machine Learning - driven insights for predicting the impact of nanoparticles on the functionality of biomolecules, Illustrated by the case of DNA Damage-Inducible Transcript 3 (CHOP) inhibitors

Mariya L. Ivanova[1,*, ORCID], Nicola Russo[1, ORCID] , Gueorgui Mihaylov[2, ORCID], Konstantin Nikolic[1, ORCID]

Author affiliations
[1]School of Computing and Engineering, University of West London, London, UK
[2]Haleon, London, UK
*Corresponding author mariya.ivanova@uwl.ac.uk

Abstract. The presented study contributes to ongoing research that aims to overcome challenges in predicting the bio-applicability of nanoparticles (NPs). The approach explored a variety of combinations of nuclear magnetic resonance (NMR) spectroscopy data derived from the Simplified molecular-input line-entry system (SMILES) notations and small biomolecule features. The resulting datasets were utilised for machine learning (ML) with scikit-learn and deep neural networks (DNN) with PyTorch. Despite the obstacles in predicting how NPs influence biomolecule functionalities, the methodology was reasoned in terms of its applicability to compounds both with and without NPs. The methodology was illustrated through a quantitative high-throughput screening (qHTS) aimed at finding DNA Damage-Inducible Transcript 3 (CHOP) inhibitors. Based on this data, the optimal ML performance was achieved by the Random Forest Classifier, which was trained on 19,184 samples and tested on 4,000, resulting in 81.1% accuracy, 83.4% precision, 77.7% recall, 80.4% F1-score, 81.1% ROC, and a five-fold cross-validation score of 0.821. Complementing the main study, two computational approaches were developed to enhance CHOP inhibitor prediction. The first identifies the most desirable/undesirable functional groups for CHOP inhibition. The second, a CID_SID ML model, achieved 90.1% accuracy in predicting whether compounds designed for other purposes possess CHOP inhibition potential.

Key words: Scikit learn, PyTorch, SMILES, NMR, CID_SID ML model.

Introduction

Nanoformulations (NFs), constructed from biomolecules and nanoparticles (NPs) into nanoscale architectures, are designed to augment biomolecule efficacy across various medical applications, such as drug delivery and tissue engineering, theragnostics, imaging, sensing, vaccine development, and medical nanodevices. This convergence of nanotechnology and medicine has spawned the field of nanomedicine. Although a relatively new field, nanomedicine has achieved substantial progress, with approximately 100 nanotherapeutics approved or under FDA review [1].

However, predicting NPs behavior, as a part of such NFs, presents significant challenges. Due to their nanoscale dimensions, NPs cannot be accurately characterized using standard light microscopy [2]. Furthermore, their dimensions approach the quantum realm and hold the potential to alter inherent properties. Brownian forces, particularly influential on sub-micron particles, complicate particle motion control. Inconsistencies observed in particle sizing using various methods emphasize that the size measurement of the NPs is not an authentic feature [3]. Moreover, human serum (HS) can alter NP`s size and surface potential due to albumin adsorption and/or fibrinogen aggregation, as shown in studies with poly(lactic-*co*-glycolic) acid (PLG) NPs in buffer saline [4]. Also, the conductance state of NPs is size-dependent. Ultrafast laser spectroscopy has revealed the transitions in gold NPs from metallic to non-metallic behaviour based on size changes [5]. Notably, NPs with similar atom counts and sizes can exhibit significantly different

toxic effects [6]. A critical concern regarding NPs prediction extends beyond well-documented issues like cytotoxicity, genotoxicity, immunogenicity, unintended organ accumulation, and long-term side effects [7, 8, 9]. To address these concerns, specialized methodologies have been developed. One such approach is the nano-quantitative structure-activity relationship (nano-QSAR), an adaptation of the well-established QSAR model used in chemistry and pharmacy [10] Nano-QSAR aims to correlate nanoparticle structure with biological activity. While promising, it currently lacks universality and requires further refinement [11]. Similarly, the structure and activity prediction network (SAPNet), designed to guide NP design by identifying structural modifications for desired properties [12], also necessitates improvement. Another QSAR-derived method utilizing simplified molecular input-line entry systems (SMILES) [13] considers the molecular structure and electrical data to predict endpoints, but its NF-specific adaptation remains incomplete, highlighting the need for further research [10]. Beyond QSAR variations, quantum mechanics has been applied to estimate nanoparticle targeted delivery efficiency [14], although the reliance on wireless electromagnetic radiation systems introduces potential sources of error. Proof-of-concept models using iron oxide NPs demonstrated the feasibility of simulating nanomaterial impacts on living organisms with machine learning (ML) [15].

The current study explored the application of ML and nuclear magnetic resonance (NMR) spectroscopy data for addressing the limitations in NP prediction. Given the cost and time required to develop medicines, which takes USD 100 million to USD 2 billion over an average period of 10 to 17 years to bring a single drug to market [16], plus the fact that overall, 9 out of 10 drug candidates that have entered clinical trials never submit for FDA approval [17], ML with its ability to predict promising- and fail- drug candidates in the early stage of their development, has the potential to prevent further investment in non-valued formulations and save a significant portion of the development cost per drug. On the other hand, the NMR technique provides information about atomic structure and their chemical environments, which information is closely related to the functionality of biomolecules [18]. Electromagnetic radiation absorption by atomic nuclei in a strong magnetic field allows for the exploitation of their magnetic properties. The subsequent relaxation of the excited nuclei, accompanied by the emission of radiation, provides a spectrum of frequencies. These frequencies, or chemical shifts, are highly sensitive to the electronic environment of the nuclei, enabling the elucidation of molecular structure and dynamics. Carbon-13 isotope (^{13}C) NMR spectroscopy identifies the chemical environment of carbon atoms within a biomolecule's carbon skeleton [19]. The proton (1H) NMR spectroscopy, on the other hand, identifies different hydrogen nuclei and their magnetic properties, revealing hydrogen bonding patterns that contribute to the understanding of intermolecular interactions [20]. The difference between both types of NMR spectra is that the former is a single peak corresponding to each unique carbon environment and is simpler than the latter, whose complexity is due to the spin-spin coupling between neighbouring protons [21]. So, considering the established dependencies of biomolecule functionalities mentioned above and the elucidative capabilities of NMR, providing information for the chemical environment of the carbon atoms building the carbon “skeleton “of the small biomolecule, it was hypothesised that an ML model trained on NMR spectroscopic data could predict the influence of NPs on biomolecule functionality, testing the NF both before and after it has been in a biological fluid such as blood plasma, cell culture media, or interstitial fluid and thus considered the influence of the “protein corona”.

The necessary NMR spectroscopy data was obtained by converting the small biomolecules` SMILES notations [13] into 13C and 1H NMR spectroscopy data using the NMRDB online tool [22]. This tool employs HOSE code (Hierarchical Organisation of Spherical Environments) methods, and the quantitative errors of the predictions can vary depending on the specific compound [23] In general, the Mean Absolute Error (MAE) of 1H NMR chemical shifts (protons) are in the range of 0.2-0.3 ppm [24], and of 13C NMR chemical shifts (carbons) are around 3ppm [25]. Even though the data used in the study were synthetic with fluctuating errors, they were sufficient to conduct and illustrate a preliminary study of the strategy concept.

In the available literature, NMR spectroscopy has been pointed as a useful addition to electron

microscopy and optical absorption spectroscopy used for characterisation of NPs, particularly for the hard–soft matter interfaces [26] and recognised as a technique capable of bridging the analytical gap between NPs in solution and solid phases [27]. NMR techniques have been employed to provide a method for elucidating the morphology and dynamics of polymer-functionalised NPs, with potential application to complex systems that form coronas around NPs [28]. Another study explored quantitative and one– and multi-dimensional NMR spectroscopy on gold NPs and developed a general method for NPs characterisation with NMR spectroscopy [29]. Overall, the direct influence of NPs on the functional activity of the small biomolecules, which they are intended to assist, requires investigation that can be supported by the NMR spectroscopy, covering Chemistry analysis of the NPs, their structural and dynamic characterisation and detection of their interactions with other molecules or materials [30].

The presented approach followed the methodology of two prior studies [31, 32] that predicted human dopamine D1 receptor antagonists and Transthyretin (TTR) transcription activators, respectively. Both studies employed ML algorithms from the scikit-learn library [33]. The ML data was derived from SMILES notations converted to ^{13}C NMR spectroscopy features by the NMRDB software. The molecular features of the small biomolecule, pre-calculated by PubChem [34], XLogP3 [35], and CACTVS [36] and provided by PubChem, have already shown their potential for ML development [37, 38] and explored in the study focused on predicting Transthyretin transcription activators [32]. These features were:

(i) Molecular weight (MW) as a sum of the mass of all constituent atoms [39].
(ii) Topological Polar Surface Area (TPSA) [40]
(iii) XLogP3-AA (XL), which is a predicted octanol-water partition coefficient [41].
(iv) Hydrogen Bond Donor Count (HBDC) in the given small biomolecule.
(v) Hydrogen Bond Acceptor Count (HDAC) in the given small biomolecule.
(vi) Rotatable Bond Count (RBC). For a bond to be rotatable, it must be a single bond, not part of a ring, and connect two atoms that are not hydrogen and are not at the end of a chain.

The current study expanded upon the methodology of the two preliminary studies by incorporating ^{1}H NMR spectroscopy data. The key difference from these preliminary studies was the use of consecutive decimal numbers, rather than natural numbers, to define chemical shift subranges for feature generation, allowing a more in-depth analysis of peak counts. Unlike the preliminary studies, which used traditional ML approaches, this study employed a PyTorch-based deep neural network (DNN) [42] with Optuna-optimized hyperparameters [43]. Python [44] and Jupyter Notebook [45] were employed as the programming language and development environment for all prediction models.

Data derived from PubChem AID 2732 bioassay [46] focused on predicting the C/EBP Homologous Protein (CHOP) inhibitors was used to demonstrate the methodology. CHOP is a crucial transcription factor in the apoptotic arm of the Unfolded Protein Response (UPR). It can be activated by the accumulation of aberrantly folded proteins that have been recognised by the cellular surveillance system and retained within the endoplasmic reticulum (ER) [47,48]. The transcription factor activates ER protein chaperones and mediates for UPR response. So, it has been hypothesised that the inhibition of CHOP could regulate the unfolded protein response to ER stress and would have a potential therapeutic application to diverse diseases [49, 50, 51], such as diabetes [52], Alzheimer's disease [53] (although, it has been reported that CHOP is not the primary contributor to tau-mediated toxicity, related to memory lost [54]), Parkinson's disease [55], haemophilia [56], lysosomal storage diseases [57], and alpha-1 antitrypsin deficiency [58].

The paper also presents two computational applications for CHOP inhibition, in addition to the main research. The first one ordered the functional groups/fragments of explored small biomolecule from the most to the least desirable with respect to the CHOP inhibition. This computational approach was based on data encoded in the chemical names derived by the International Union of Pure and Applied Chemistry (IUPAC) nomenclature, which by itself ensure unique names for any chemical compound. The methodology has been developed

and demonstrated with a case study on Tyrosyl-DNA phosphodiesterase 1 (TDP1) inhibitors [59]. Through developing this approach with CHOP-related data, the results serve researchers interested in CHOP inhibition and explore the applicability of the methodology for different data than TDP1. The second computational approach developed beyond the main study was the CID_SID ML model that can predict CHOP inhibitors, using only PubChem identifiers, i.e. PubChem CID and PubChem SID. This approach enables the assessment of small biomolecules initially intended for other targets for their CHOP inhibition capability. Since, generally, the identifiers are not used for ML training and testing, the CID_SID ML model is unconventional. Despite this, its development was meaningful because PubChem's method considers structural and similarity data when generating their identifiers. [60]. The CID_SID ML model has already been developed and is available in the relevant study [61].

Methodology

The methodology is illustrated in Figure 1 and Figure 2. The columns with CIDs, SMILES notations and activity labels of small biomolecules were retrieved from the PubChem AID 2732 dataset [46]. Since the study was focused on classification ML models, the severe imbalance between the inactive and active small biomolecules was handled. For this purpose, the inactive samples were reduced by keeping only the compounds considered as well in PubChem AID 1996 bioassay focused on small biomolecule solubility [62]. After shuffling the reminded inactive compounds, each second sample was selected and kept. Thus, the inactive compounds were reduced to some extent and combined with all active compounds from the PubChem AID 2732 dataset [46].

From the resulting dataset, the following datasets were created:

(i) A dataset, which included only CIDs and SMILES notations, was formulated to enable the acquisition of NMR spectroscopic data through the NMRDB software.

(ii) A dataset containing only SMILES notations and the activity labels of the small molecule was used later for labelling the spectroscopic data

(iii) A dataset containing only CIDs, which was used as a list for downloading the molecule features listed above of the small molecules.

Once the spectroscopic chemical shifts were obtained for 1H NMR and 13C NMR spectroscopy data, two types of datasets were generated. In the first type, each pick along the chemical shifts scale was counted within subranges defined by consecutive integers and called concise. The subranges in the second type were defined by consecutive decimal numbers and called extensive. All subranges formed the newly generated features of the data frames, which contained the number of picks along the chemical shifts scale. The four initial datasets, comprising both concise and extensive 1H NMR and 13C NMR spectroscopy data, were combined, as shown in Table 1, to yield eight resulting datasets. Each of these eight datasets was used further for ML with the classifiers: Decision Tree, Random Forest, Support Vector and Gradient Boosting software interpreted by scikit-learn ML library [35]. ML was conducted based on the best practices recommended in the literature [63, 64]. For that purpose, an equal number of samples for each class were extracted to ensure that there would not be bias towards the majority class that would lead to misleading accuracy. The remaining samples were then balanced, increasing the number of minor classes with randomly selected and repeated samples of this class until the number of samples in the minor class was equal to the number of samples in the major class. The ML models were then conducted with each of these eight datasets. The ML metrics were compared, and the most suitable dataset and an ML classifier for this case study were selected and scrutinised for overfitting tracing the deviation between the training and testing accuracy to be lower than 5%.

Further, the above-described molecular features were integrated into the dataset used by the optimal ML model. Following this, ML analyses were conducted, comparing the results with and without PCA application [65], which application reduced the number of features, and the results obtained using only the molecular features. Using the expanded dataset with molecular features, a PyTorch DNN was developed, with hyperparameters optimised by Optuna [43] and scrutinised for overfitting. DNN was trained ten times, and the mean accuracy was compared to the optimal scikit-learn ML model.

The supplementary computational approaches followed the methodologies of the relevant studies. About this one using data encoded in the IUPAC names [59], the IUPAC names of the small biomolecule from the PubChem AID 2732 bioassay` dataset [46] were parsed into strings of four or more letters, and then used in two ways: to develop an ML model and to extract the most desirable and non-desirable functional groups/fragment regarding the CHOP inhibition. For the second ML approach, the CID_SID ML model, the CIDs, SIDs and targets from the PubChem AID 2732 bioassay` dataset [46] were extracted and used for ML [61]. For more details about the methodologies, please refer to the relevant studies [59, 61].

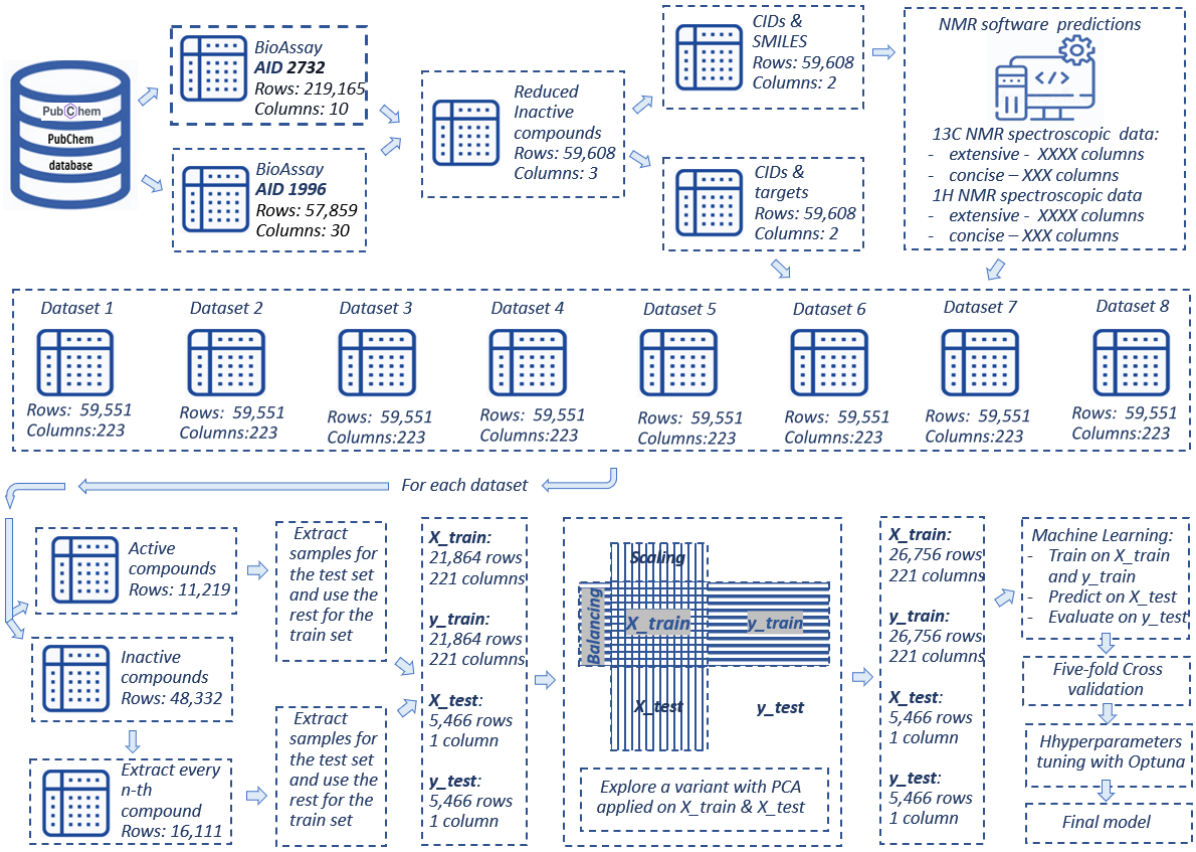

Figure 1 Methodology of ML with eight datasets

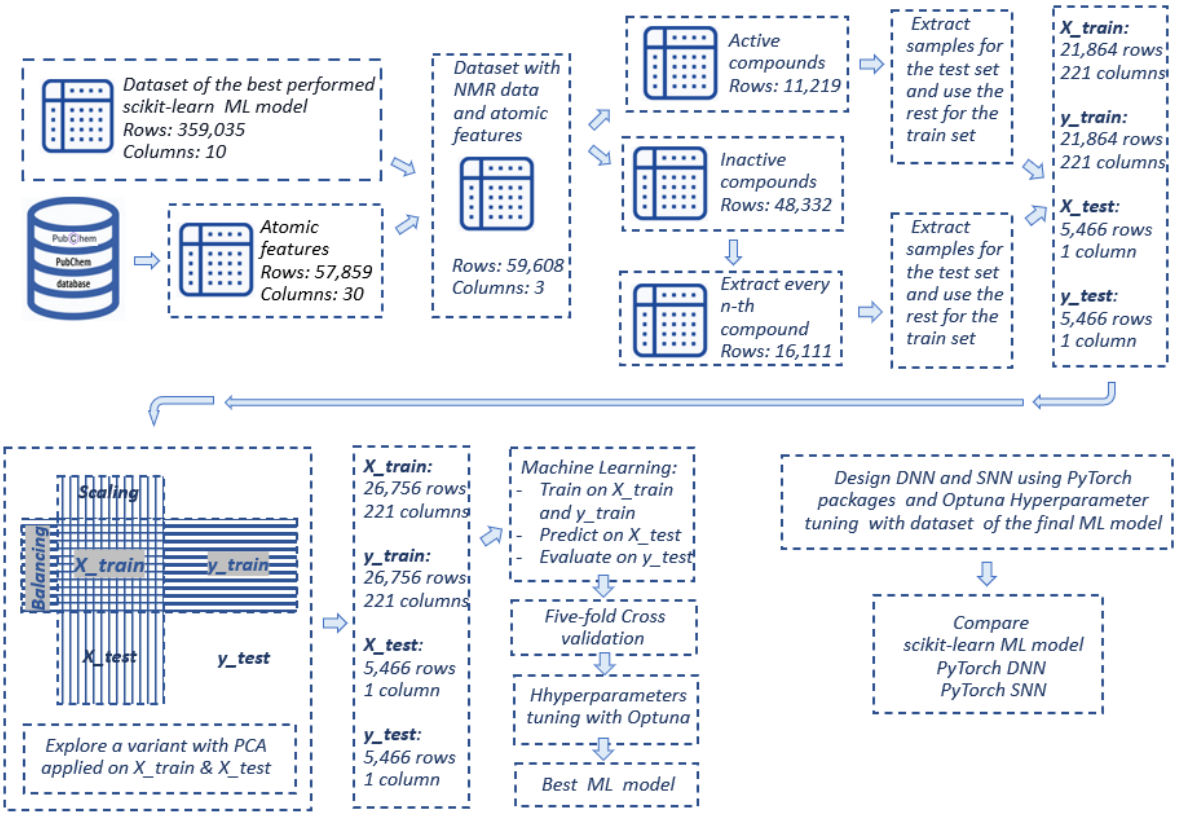

Figure 2 ML performed with the dataset of the optimal ML model and molecular features

Results and discussion

By identifying the overlap between 210,922 inactive compounds from PubChem AID 2732 and 57,859 samples from PubChem AID 1996, 24,185 inactive samples were retained. The number of these compounds was reduced subsequently, keeping every second sample, which decreased them to 12,085. Combining these 12,085 remaining inactive samples with the 8,224 active compounds from PubChem AID 2732 resulted in a dataset of 20,309 samples. The SMILES notations from this dataset were used by the NMRDB software, and NMR spectroscopy data was obtained as follows:

(i) 220 columns with counts of chemical shifts of 13C NMR spectra in subranges defined by consequent natural numbers; the dataset abbreviation is 13C_c.

(ii) 12 columns with counts of chemical shifts of 1H NMR spectra in subranges defined by consequent natural numbers; the dataset abbreviation is 1H_c.

(iii) 1,828 columns with counts of chemical shifts of 13C NMR spectra in subranges defined by consequent natural numbers; the dataset abbreviation is 13_e.

(iv) 103 columns with counts of chemical shifts of 1H NMR spectra in subranges defined by consequent natural numbers; the dataset abbreviation is 1H_e.

These four datasets, containing NMR spectroscopy data, were combined as detailed in Table 1, resulting in eight datasets subsequently used for ML. To ensure reliable metric results, 2,100 samples were randomly selected from each dataset, resulting in a total of 4,200 samples used for testing the ML models. The imbalance in the remaining compounds was handled using Random Over Sampling (ROS), a technique replicating random minority class samples until both classes contained an equal number of samples. This process yielded a balanced dataset of 16,109 samples, which was used for training the ML models. Since there was a risk of the minority and majority classes not being well-separated, which could lead to synthetic samples blurring the decision boundary, ROS was chosen over Synthetic Minority Over-sampling Technique (SMOTE). Additionally, ROS was preferred over

Random Undersampling (RUS) to avoid decreasing the overall data volume, which is crucial for ML model performance.

Table 1 summarises the optimal ML models' accuracy and five-fold cross-validation scores across the eight datasets. The performance of the ML models, when incorporating 13C NMR data, ranged from 76.4% (Dataset 6) to 79.3% (Dataset 2), a difference that was not statistically significant. Furthermore, including 1H NMR data did not improve model accuracy; in fact, a slight decrease was observed. Models trained solely on 1H NMR data yielded the lowest accuracies across the eight datasets, with results of 67.7% (Dataset 3) and 69.5% (Dataset 4). The large gap between single result accuracy and five-fold cross-validation scores indicated potential overfitting. Due to this pattern across all datasets, only the optimal ML model (SVC, Dataset 2) was scrutinized for overfitting.

Initially, SVC based on Dataset 2 (extensive 13 C NMR spectroscopy data) achieved 79.3% accuracy, 82.6% precision, 74.1% recall, 78.2% F1-score, 79.3% ROC (Table ESM 3), and a 0.835 five-fold cross-validation score (standard deviation ±0.002) (Table ESM4) followed closely by SVC based on Dataset 8 (extensive 13 C NMR and extensive 1 H NMR spectroscopy data) 79.2% accuracy, 82.4% precision, 74.2% recall, 78.1% F1-score, 79.2% ROC (Table ESM15) and a 0.844 five-fold cross-validation score (standard deviation ±0.002) (Table ESM16). The SVC variant trained on Dataset 2 was selected to minimize training and testing time due to its lower feature dimensionality. On the other hand, considering both variants, five-fold cross-validation revealed that RFC consistently had the highest cross-validation score, thus ranking it first. However, the substantial difference in RFC performance between a single evaluation and five-fold cross-validation suggests potential overfitting (Table ESM 3, Table ESM4, Table ESM15, Table ESM16). Initial analysis showed no significant overfitting in the SVC model (with default hyperparameter based on Dataset 2) with training and testing accuracies of 0.846 and 0.793, respectively. Optuna's five-trial hyperparameter optimization (C=218090.43, gamma=0.0518) resulted in no significant accuracy gain. However, it increased the training-testing accuracy delta (0.879 vs. 0.796), implying a higher propensity for overfitting.

The inclusion of molecular features in Dataset 2 altered its dimensions, increasing the number of features to 1,934 and reducing the number of samples to 19,501. The optimal ML model performed with this dataset was RFC achieving 83% accuracy, 88% precision, 76.4% recall, 81.8% F1-score, 83% ROC (Table ESM17), and a 0.84 five-fold cross-validation score with 0.006 standard deviations (Table ESM18). The overfitting assessment indicated that max_depth values exceeding 13, where the training and testing accuracy were 84.6% vs. 79.8%, respectively, resulted in a training-testing accuracy deviation greater than 5%, implying a potential for overfitting (Figure ESM 1). The hyperparameter tuning of RFC with Optuna improved the performance of the ML model to 81.1% accuracy, 83.4% precision, 77.7% recall, 80.4% F1-score, 81.1% ROC and 0.821 five-fold cross-validation score. Overfitting (below max_depth=9) was not indicated by the examination (Figure 3). The hyperparameter values suggested by Optuna were:

(i) max_depth=9, define the level the tree can have.
(ii) n_estimators=494, shows the number of trees in the forest.
(iii) min_samples_split=2, the minimum number of samples required to split an internal node.
(iv) min_samples_leaf=6, the minimum number of samples required to be at a leaf node.
(v) max_features=None, i.e. max_features=n_features.
(vi) criterion='entropy', measuring the quality of a split.

The learning curve of the ML model is plotted in Figure 4, the confusion matric in Figure 5, the AUC in Figure 6, and the classification report in Table 2.

Dataset 2, with added molecular features, was used for ML after PCA reduced its dimensionality to the optimal eight components. However, this did not yield performance gains compared to the model without PCA (Table ESM19, Table ESM20). An ML variant using only molecular features was explored, but it did not outperform RFC with molecular-feature-integrated Dataset 2 as well (Table ESM21, Table ESM22).

The Optuna-optimized DNN [43] comprised two hidden layers with 113 and 104 units, respectively.

Dropout rates of 0.33165 and 0.3692 were applied after each layer. Utilizing the RMSprop optimizer with a learning rate of 0.001329, the model achieved 82.34% accuracy. However, a significant discrepancy between the final training loss (0.07) and validation loss (2.5), as depicted in Figure ESM2, suggested potential overfitting.

Table 1. Content of the eight datasets which were result of the combination of NMR spectroscopy data; Accuracy of the best ML model with the given dataset before to be scrutinized for overfitting; five-fold cross-validation score of the respected ML model; references to the tables in Electronic Supplementary material (ESM) with full set of ML metrics. Datasets contain data of carbon-13 isotope concise (13C_c), carbon-13 isotope extensive (13C_e), proton isotope concise (1H_c), proton isotope extensive (1H_c),

Dataset	13C_c	13C_e	1H_c	1H_e	Accuracy	c-v score	Table
1	✓	-	-	-	77.9%	0.864	ESM1, ESM2
2	-	✓	-	-	79.3%	0.864	ESM3, ESM4
3	-	-	✓	-	67.7%	0.769	ESM5, ESM6
4	-	-	-	✓	69.5%	0.821	ESM7, ESM8
5	✓	-	✓	-	77.8%	0.861	ESM9, ESM10
6	-	✓	✓	-	76.4%	0.872	ESM11, ESM12
7	✓	-	-	✓	77.8%	0.861	ESM13, ESM14
8	-	✓	-	✓	79.2%	0.861	ESM15, ESM16
Dataset 2 & molecule features					83.0%	0.840	ESM17, ESM18
Dataset 2 & molecule features; feature reduction with PCA					82.8%	0.902	ESM19, ESM20
Dataset only with molecule features					80.7%	0.887	ESM21, ESM22
PyTorch DNN with Dataset 2 & molecule features					81.3%	-	Figure ESM2

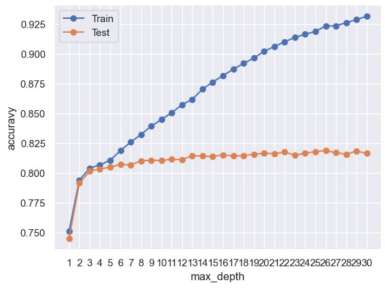

Figure 3 Scrutinising for overfitting of the Optuna-hyperparameter tuned RFC
based on Dataset 2 and molecular feature

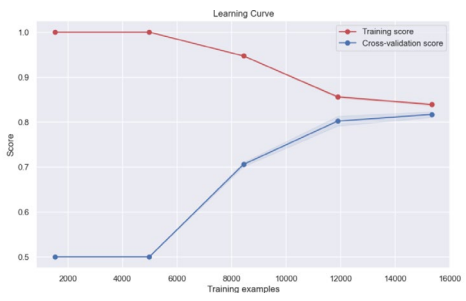

Figure 4 Learning curve of the Optuna-hyperparameter tuned RFC
based on Dataset 2 and molecular features

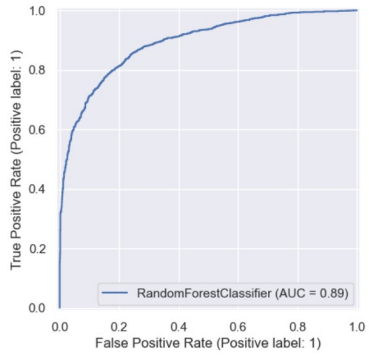

Figure 5. ROC of the Optuna-hyperparameter tuned RFC based on Dataset 2 and molecular features

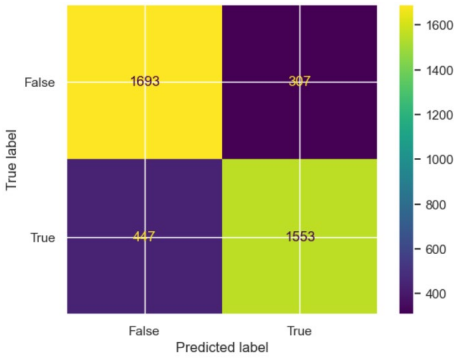

Figure 6. Confusion matrix of the Optuna-hyperparameter tuned RFC based on Dataset 2 and molecular features

Table 2. Classification report of the Optuna-hyperparameter tuned RFC based on Dataset 2 and molecular features

	precision	recall	F1-score	support
Active (target 1)	0.79	0.85	0.82	2000
Inactive (target 0)	0.83	0.78	0.80	2000
accuracy			0.81	4000
macro avg	0.81	0.81	0.81	4000
Weighted avg	0.81	0.81	0.81	4000

The Optuna-tuned RFC achieving 68.9% accuracy, 71.6% precision, 62.5% recall, 66.7% F1-score, 68.8% ROC AUC, and a five-fold cross-validation score of 0.724% (standard deviation ±0.0076). The optimal hyperparameters, determined by Optuna were min_samples_split=10, min_samples_leaf=1, max_features='log2', criterion='gini', and max_depth=15, which was set to prevent overfitting (train/test accuracy deviation < 10%). Figure ESM 3 illustrates the scrutinising for overfitting of the model with default hyperparameters, where the overfitting started at max_depth = 19, where the train accuracy was 66.2%; Figure ESM 4 is a plot of the scrutinising for overfitting of the model hyperparameter tuned by Optuna, which was the final model; Figure ESM 5 confusion matrix and Table ESM 23 the classification report of the final model.

The top 24 functional groups (Figure ESM 6, Figure ESM 7), as ranked by the feature importance and permutation importance algorithms, showed high similarity. The difference between them was three groups per list, namely amino, pyramidin, methyl, sulfamoyl, carbamothioyl, trichloro. The result of combination of two feature importance lists keeping only the unique functional groups/fragments was reordered according to the relevant proportion of the active cases, so with the highest value was sulfamoyl with 69 active and 13 inactive cases and relative proportion of 5.31 and the lowest value relative proportion was hold by carbohydrazide with 0.1 relative proportion of the active cases. The full reordered list based on relative proportion of the active cases is available on GitHub [66]. It should be noted that sulfamoyl should participate in the small biomolecule composition as an independent functional group, not in a pact with other groups, for example like sulfamoylbenzoicacid because it turned out that sulfamoyl name participated in 26 features and formed dataset of 19,499 compounds, where 11,593 are inactive and 7,906 are active, so just the presence of sulfamoyl in the IUPAC name is not enough to conclude that there is a high probability that the compound is a CHOP inhibitor. The sulfamoyl functional group should not be in a composition with other functional groups.

The entire dataset of the bioassay PubChem AID 2732 [46] was used for generation of the lists with functional groups/fragments participate only in active cases (highly desirable) or inactive cases

(highly undesirable). As it was mentioned above, the dataset contained 218,583 samples 8,224 of which were labelled as active and 209,952 as inactive. So, the list with highly desirable functional groups/fragments was with 267 elements, starting with hexahydroazuleno and dioxonaphthalene as the most desirable each one of them with five active cases and no one inactive, followed by ynamide, oxirane, pyrazolidine each of them with four active cases The full list is available on GitHub [67]. It is worth mentioning that the hypothesis proposed in the parent study of this approach [59] suggested that when the tested compound contains a functional group from one of these two lists and its (of the tested compound) 13CNMR spectroscopy data is similar to the compound source of this functional group/fragment, there is a high probability that the tested compound is a CHOP inhibitor. Given the nature of 13C NMR spectroscopy, the hypothesis was expected to be applicable in the presence of NPs (i.e. for NFs) as well. In this case the NFs will be processed in the same manner explained in the main study, i.e. way explained in the main study, i.e. NFs to be stood in a complex biological fluid in order to provoke NPs aggregation and then to process NMR spectroscopy whose data will be used for comparison.

Regarding the most undesirable functional groups/fragments, i.e. whose compounds participate only in inactive cases, the descendent order contained 7,155 rows on the top of which was methyl with 25,955 inactive cases followed by carboxamide with 18,520 and methoxyphenyl with 13,153. The list with the first 3000 rows is available on GitHub [68].

Regarding CID_SID ML model that check if a compound, designed initially for other purpose different than CHOP inhibition is a CHOP inhibitor, the RFC obtained 90.1% accuracy, 98.3% precision, 81.7% recall, 89.2% F1, 90.1% ROC, five-fold cross-validation score of 0.943 with standard deviation of ±0.00075. For more details, please refer to the original study [61]

Conclusion

An ML methodology was developed, leveraging molecular features and 1H and 13C NMR spectroscopy data derived from SMILES notations. It was hypothesised that the approach could be applicable in case of predictions of a NF functionality due to the capability of 13C NMR to provide information regarding the chemical environment of the carbons building the carbon skeleton of small biomolecules. So, the presence of nanoparticle with or without protein corona would affect this environment and respectively detected by the 13C NMR spectroscopy. This information in turn can be used by ML to find a pattern amongst data that can predict the functionality of the compound (NF) accommodating the NP. The methodology was demonstrated using CHOP inhibitors as a case study, but its applicability to other bioassays is anticipated. A key innovation of this research, within a broader investigation, was the refined segmentation of chemical shift ranges, which increased feature dimensionality and slightly enhanced ML model performance. Additionally, a PyTorch DNN was designed and optimized using Optuna. The optimal solution was RFC that scrutinised for overfitting to confirm its robustness. Although data used for the development of the ML model was synthetic, the NMR chemical shift of NF that will be tested must be real spectroscopy data. The complementary computational approaches provide the researchers interested in CHOP inhibition with insights related to drug discovery and a side effect.

Scientific contributions

- Proposed a new concept to address the challenges in predicting the influence of NPs (NPs) based on 13C NMR spectroscopy.
- Developed an ML model for predicting CHOP inhibition based on SMILES notations.
- Sulfamoyl was identified as having potential to contribute to CHOP inhibition
- Hexahydroazuleno and dioxonaphthalene were identified as a highly desirable functional group for CHOP inhibition.
- Methyl was found to be a highly undesirable functional group for CHOP inhibition
- Developed CID_SID ML model that leverages PubChem CID and SID to predict CHOP inhibition as a potential side effect for existing chemical compounds.

Author Contributions

MLI, NR, GM and KN conceptualized the project and designed the methodology. MLI and NR wrote the code. MLI, NR and GM processed the data. KN supervised the project. All authors were involved with the writing of the paper.

Acknowledge

MLI thanks the UWL Vice-Chancellor's Scholarship Scheme for their generous support. We sincerely thank NIH for providing access to their PubChem database. Article is dedicated to Luben Ivanov

Data and Code Availability Statement

The raw data used in the study is available through the PubChem portal:
https://pubchem.ncbi.nlm.nih.gov/

The code generated during the research is available on GitHub:
https://github.com/articlesmli/13C_NMR_ML_model_CHOP.git

Conflicts of Interest

The authors declare no conflict of interest.

Authors` biographies

Mariya L. Ivanova (ORCID) is a Ph.D. candidate in Computer Science at the University of West London, UK. She holds two master's degrees: one in Electrical Engineering from the Technical University of Sofia, Bulgaria, and another in Software Engineering from the University of West London. Previously, she worked as a Machine Learning Engineer at 885Ltd in London, UK.

Dr. Nicola Russo (ORCID) holds a Ph.D. in Artificial Intelligence from the University of West London, UK, and a master's degree in Security of Software Systems from the University of Naples "Parthenope" in Italy. He is currently an HPL Lecturer in Computer Science at the University of West London. His previous roles include a Research Assistant position at the University of West London and an HPL Lecturer at Middlesex University in London, UK.

Dr. Gueorgui Mihaylov (ORCID) completed a MSc in Theoretical Physics (2004) and a PhD in Mathematics (2008) with a thesis in differential geometry at the University of Milan. Currently GM is a Principal Data Scientist (AI/ML Director) at Haleon-former GSK Consumer Healthcare, Visiting Research fellow at the Department of Mathematics of King's College London and a member of the EPSRC Strategic Advisory Team for Mathematical Sciences. In the past GM had a series of postdoc and other research and teaching positions at University of Turin, Polytechnic University of Turin and Polytechnic University of Milan, a visiting professorship in applied mathematics at the Silesian University of Technology etc. In 2017, GM joined Royal Mail as a Senior and later as a Principal Data Scientist. GM has been contributing and leading multiple industrial and academic research projects with a strong component of mathematical modelling, optimisation, and optimal control on complex (industrial) systems, industrial applications of ML and AI, geometry and topological methods in data science, manifold learning etc.

Prof. Konstantin Nikolic (ORCID) (Member, IEEE) received the master's degree in applied physics from the University of Belgrade, Serbia, and the Ph.D. degree in physics from Imperial College London, London, UK. He is currently a Professor of Computer Science: AI, machine learning, and data science with the School of Computing and Engineering, University of West London, and a Visiting Professor with the Department of Electrical and Electronic Engineering, Imperial College, London. Previously, he was Corrigan Research Fellow, then a Senior Research, Fellow and the Principal Investigator at the Institute of Biomedical Engineering and the Department of Electrical and Electronic Engineering, Imperial College London, from June 2006 to February 2020. Before that, he was a Senior Research Associate with UCL, and an Assistant Professor and then an Associate Professor with the Faculty of Electrical Engineering, University of Belgrade. He leads NeuroAI Group, which develops methods and computational tools for understanding, modeling, and simulating various biological and physiological processes, and their applications in AI/ML, bio-inspired electronic systems, and diagnostics. He is a member of the IEEE CAS Technical Committee and the Royal Society Neural Interfaces Steering Group.

References

[1] X. Shan et al., “Current approaches of nanomedicines in the market and various stage of clinical translation”, *Acta Pharm. Sin. B*, vol.12, no. 7, pp. 3028–3048, 2022, doi: 10.3389/fped.2024.1396408

[2] A. Ponta, “Considerations for Drug Products that Contain Nanomaterials”, *FDA*, May 2024 [online] https://www.fda.gov/drugs/cder-small-business-industry-assistance-sbia/considerations-drug-products-contain-nanomaterials#:~:text=FDA%20recently%20released%20the%20guidance,to%20conventional%20manufacture%20or%20storage (accessed 10 January 2025)

[3] K. Eitel, G. Bryant and H.J. Schöpe, “A Hitchhiker’s Guide to Particle Sizing Techniques”, *Langmuir,* vol. 36, no. 35, pp. 10307-10320, 2020, doi: 10.1021/acs.langmuir.0c00709

[4] C. Fornaguera, G. Calderó, M. Mitjans, M.P. Vinardell, C. Solansa and C. Vauthierc, “Interactions of PLGA nanoparticles with blood components: protein adsorption, coagulation, activation of the complement system and hemolysis studies”, *Nonoscale,* vol. 14, no. 7, pp. 6045-6058, 2015, doi: 10.1039/C5NR00733J

[5] M. Zhou et al., “Evolution from the plasmon to exciton state in ligand-protected atomically precise gold nanoparticles”*, Nat. Commun.,* vol. 7, no. 10, p. 13240, 2016, doi: 10.1038/ncomms13240

[6] A.V. Singh et al., “Artificial Intelligence and Machine Learning in Computational Nanotoxicology: Unlocking and Empowering Nanomedicine”, *Adv. Healthc. Mater.*, vol.9, no. 17, p. 1901862, 2020, doi: 10.1002/adhm.201901862

[7] H. Salehi, M. Etemadi and P. Kazemi, “Current Trends and Challenges in Pharmacoeconomic Aspects of Nanocarriers as Drug Delivery Systems for Cancer Treatment.”, *Front. Pharmacol*., vol. 12, 763403, 2021, doi: 10.2147/ijn.s323831

[8] T. L. Moore et al, “Nanoparticle colloidal stability in cell culture media and impact on cellular interactions”, *Chem. Soc. Rev*., vol. 44, no. 17, 6287-6305, 2015, doi: 10.1039/C4CS00487F

[9] D. T. Savage, J. Z. Hilt and T. D. Dziubla, “In Vitro Methods for Assessing Nanoparticle Toxicity”, *Methods Mol Biol*., vol. 1894, pp. 1-29., 2019 doi: 10.1007/978-1-4939-8916-4_1

[10] I. Furxhi, F. Murphy, M. Mullins, A. Arvanitis and C.A. Poland, “Practices and Trends of Machine Learning Application in Nanotoxicology”, *Nanomaterials* (Basel, Switzerland), vol. 10, no. 1, p.116, 2020, doi: 10.3390/nano10010116

[11] Q Qi and Z. Wang, “Integrating machine learning and nano-QSAR models to predict the oxidative stress potential caused by single and mixed carbon nanomaterials in algal cells”, *Environmental Toxicology and Chemistry*, 2025;, vgae049, doi: 10.1093/etojnl/vgae049

[12] A. Rybinska-Fryca, A. Mikolajczyk and T. Puzyn, “Structure–activity prediction networks (SAPNets): a step beyond Nano-QSAR for effective implementation of the safe-by-design concept”, *Nanoscale*, vol. 40, no. 10, pp. 20669-20676, 2020, doi: 10.1039/D0NR05220E

[13] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules”, *J. Chem. Inf. Model.,* vol. 28, no. 1, pp. 31-36, 1988, doi: 10.1021/ci00057a005

[14]H. Nieto-Chaupis, "Computational Simulation of Artificial Nanoparticles Paths," *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Laguna Hills, CA, USA, 2020, pp. 193-197, doi: 10.1109/AIKE48582.2020.00045

[15] B. Rudrapath et al., "Predicting Nano Particle by Biological behavior-A Machine Learning Approach", *NanoWorld J*, 2023, doi: 10.17756/nwj.2023-s3-169

[16] S. K. Niazi and Z. Mariam, "Artificial intelligence in drug development: reshaping the therapeutic landscape", *Ther. Adv. Drug Saf.,* vol.16, 2025, doi: 10.1177/204209862513217

[17] H. Guo, X. Xing, Y. Zhou, W. Jiang, X. Chen, T. Wang, et al. "A Survey of Large Language Model for Drug Research and Development.", *IEEE Access*., vol. 13, pp. 51110-51129, 2025, doi: 10.1080/14656566.2022.2161366

[18] G.R. Fulmer et al., "NMR Chemical Shifts of Trace Impurities: Common Laboratory Solvents, Organics, and Gases in Deuterated Solvents Relevant to the Organometallic Chemist", *Organomet.*, vol. 29, no. 9, pp. 2176-2179, 2010, doi: 10.1021/om100106e

[19] J. Abramson et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3", *Nature*, vol. 630, pp. 493–500, 2024, doi: 10.1038/s41586-024-07487-w

[20] M. Milic et al. "NMR Quantification of Hydrogen-Bond-Accepting Ability for Organic Molecules", *J. Org. Chem.,* vol. 86, no. 9, pp. 6031-6043, 2021, doi: 10.1021/acs.joc.0c02876

[21] P. H. Kowalski, A. Krzemińska, K. Pernal, E. Pastorczak, "Dispersion Interactions between Molecules in and out of Equilibrium Geometry: Visualization and Analysis, *J. Phys. Chem. A*, vol. 126, no. 7, pp. 1312-1319, 2022 doi: 10.1021/acs.jpca.2c00004

[22] *N*MRDB Tools for NMR spectroscopists; Predict 13C NMR. https://nmrdb.org/13c/index.shtml?v=v2.138.0 (accessed 2025-10-14)

[23] T. Sajed et al., "Accurate Prediction of 1H NMR Chemical Shifts of Small Molecules Using Machine Learning." *Metabolites, vol.*14, no. 5, p. 290, 2024, doi: 10.3390/metabo14050290

[24] S. Kuhn S. and Johnson S.R. "Stereo-aware extension of HOSE codes.", *ACS Omega*, vol.4, pp. 7323–7329, 2019, doi: 10.1021/acsomega.9b00488

[25] C. Han, D. Zhang, S. Xia, and Y. Zhang, "Accurate Prediction of NMR Chemical Shifts: Integrating DFT Calculations with Three-Dimensional Graph Neural Networks.", *J. Chem. Theory Comput.,* vol. 20, no. 12, pp. 5250-5258, 2024, doi: 10.1021/acs.jctc.4c00422

[26] L. E. Marbella and J. E. Millstone, "NMR Techniques for Noble Metal Nanoparticles" *Chem. Mater*, vol. 27, no. 8, pp. 2721-2739, 2015, doi: 10.1021/cm504809c

[27] F. De Biasi, F. Mancin and F. Rastrelli, "Nanoparticle-assisted NMR spectroscopy: A chemosensing perspective", *Prog. Nucl. Magn. Reson. Spectrosc*., vol. 117, pp. 70-88, 2020, doi: 10.1016/j.pnmrs.2019.12.001

[28] Y. Zhang, C.G. Fry, J.A. Pedersen and R.J. Hamers, "Dynamics and Morphology of Nanoparticle-Linked Polymers Elucidated by Nuclear Magnetic Resonance", *Anal. Chem.,* vol. 89, no. 22, pp. 12399-12407, 2017, doi: 10.1021/acs.analchem.7b03489

[29] C. Guo and J. L. Yarger, “Characterizing gold nanoparticles by NMR spectroscopy. *Magn Reson Chem*., vol.56, no. 11, pp.1074–1082, 2018, doi: 10.1002/mrc.4753

[30] I. C. Felli and R. Pierattelli, “13C Direct Detected NMR for Challenging Systems”*, Chem. Rev.,* vol. 122, no.10, pp. 9468-9496, 2022, doi: 10.1021/acs.chemrev.1c00871

[31] M.L. Ivanova, N. Russo and K. Nikolic, “Leveraging 13C NMR spectrum data derived from SMILES for machine learning-based prediction of a small molecule functionality: a case study on human Dopamine D1 receptor antagonists”, *ArXiv,* 2025, 10.48550/arXiv.2501.14044

[32] M.L. Ivanova, N. Russo and K. Nikolic, “Comparative Analysis of Computational Approaches for Predicting Transthyretin Transcription Activators and Human Dopamine D1 Receptor Antagonists”, *ArXiv,* 2025, 10.48550/arXiv.2506.01137

[33] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *JMLR*, vol. 12, pp. 2825-2830, 2011, https://scikit-learn.org/stable/about.html

[34] PubChem, “Explore Chemistry”, *National Institutes of Health.* https://pubchem.ncbi.nlm.nih.gov/ Accessed 20 February 2025

[35] Xlogp3, http://www.sioc-ccbg.ac.cn/skins/ccbgwebsite/software/xlogp3/ Accessed 4 Jan 2025

[36] W.D. Ihlenfelldt, Y. Takahahashi, H. Abe and S. Sasaki, “Enhanced CACTVS browser of the open NCI database”, *J. Chem. Inf. Comput. Sci.*, vol. *42*, pp. 46-57, 2002, doi: 10.1021/ci010056s

[37] M. L. Ivanova, N. Russo, N. Djaid and K. Nikolic, “Application of machine learning for predicting G9a inhibitors”, *Digital Discovery*, vol. 3, no.10, pp. 2010-2018, 2024, doi: 10.1039/D4DD00101J

[38] M.L. Ivanova, N. Russo and K. Nikolic, “Targeting Neurodegeneration: Three Machine Learning Methods for Discovering G9a Inhibitors Using PubChem and Scikit-Learn”, *ArXiv,* 2025, doi:10.48550/arXiv.2503.16214

[39] A.M.H. van der Veen, J. Meija, A. Possolo, Antonio and D.B. Hibbert, "Interpretation and use of standard atomic weights (IUPAC Technical Report)", *Pure Appl. Chem.,* vol. 93, no. 5, 2021, pp. 629-646, doi: 10.1515/pac-2017-1002

[40] P. Ertl, B. Rohde and P. Selzer, “Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties”, *J Med Chem*, vol. 43, no. 20, pp. 3714-7, 2000, doi: 10.1021/jm000942e

[41] T. Cheng et al., “Computation of octanol-water partition coefficients by guiding an additive model with knowledge”, *J. Chem. Inf. Model.*, vol.47, no. 6, pp. 2140-8, 2007, doi: 10.1021/ci700257y

[42] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library”, In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019. [online] https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (accessed 10 January 2025)

[43] T. Akiba et al. “Optuna: A next generation hyperparameter optimization framework”, *ArXiv*, 2019, doi: 10.48550/arXiv.1907.10902

[44] G. Van Rossum, J. F. Drake, Python (Version 3.12.3), Centrum voor Wiskunde en Informatica Amsterdam, 1995. [Computer software]. Retrieved from https://www.python.org/

[45] Jupyter, *Home page*. *Jupyter.* 2024 https://jupyter.org/ (accessed 4 Jan 2025)

[46] National Center for Biotechnology Information, "PubChem Bioassay Record for AID 2732, HTS for small molecule inhibitors of CHOP to regulate the unfolded protein response to ER stress, Source: Emory University Molecular Libraries Screening Center", *PubChem* [online] https://pubchem.ncbi.nlm.nih.gov/bioassay/2732 (accessed 30 March 2025)

[47] H. Hu, M. Tian, C. Ding, S. Yu, "The C/EBP Homologous Protein (CHOP) Transcription Factor Functions in Endoplasmic Reticulum Stress-Induced Apoptosis and Microbial Infection", Front *Immunol*., vol. 9, p. 3083. doi: 10.3389/fimmu.2018.03083

[48] A. Fribley, K. Zhang, R.J. Kaufman," Regulation of apoptosis by the unfolded protein response", Methods Mol Biol., vol. 559: pp.191-204, 2009, doi: 10.1007/978-1-60327-017-5_14

[49] A. Stilkerich et al., "Cell Homeostasis or Cell Death—The Balancing Act Between Autophagy and Apoptosis Caused by Steatosis-Induced Endoplasmic Reticulum (ER) Stress", *Cells*., vol.14, no.6, p. 449, 2025, doi: 10.3390/cells14060449

[50] W. Zhang et al., "Endoplasmic reticulum stress—a key guardian in cancer", *Cell Death Discov.*, vol.10, p. 343, 2024, doi: 10.1038/s41420-024-02110-3

[51] Y. Yang et al., "Endoplasmic reticulum stress and the unfolded protein response: emerging regulators in progression of traumatic brain injury", *Cell Death Dis.* vol.15, p. 156, 2024, doi: 10.1038/s41419-024-06515-x

[52] Z. He et al., "The role of endoplasmic reticulum stress in type 2 diabetes mellitus mechanisms and impact on islet function", PeerJ., vol.13, p. e19192, 2025, doi: 10.7717/peerj.19192

[53] P. M. Sleiman et al, "Trans-ethnic Genomic Informed Risk Assessment for Alzheimer's disease: An International Hundred K+ Cohorts Consortium Study." *Alzheimers Dement.* Online July 14, 2023, doi: 10.1002/alz.13378

[54] M. Criado-Marrero, L.J. Blair, "CHOP is not a main contributor to tau-mediated toxicity", *Alzheimer's Dement*., vol. 17, p. e058717, 2021, doi: 10.1002/alz.058717

[55] P. Aimé et al., "The drug adaptaquin blocks ATF4/CHOP-dependent pro-death Trib3 induction and protects in cellular and mouse models of Parkinson's disease", *Neurobiol Dis.*, vol.136, p. 104725. 2020, doi: 10.1016/j.nbd.2019.104725

[56] A.R. Sternberg *et al.* "Pre-clinical evaluation of an enhanced-function factor VIII variant for durable haemophilia A gene therapy in male mice", *Nat Commun.,* vol. 15, p. 7193, 2024, doi: 10.1038/s41467-024-51296-8

[57] E.A. Liu, A.P. Lieberman, "The intersection of lysosomal and endoplasmic reticulum calcium with autophagy defects in lysosomal diseases", *Neurosci. Lett*., vol. 697, pp. 10-16, 2019, doi: 10.1016/j.neulet.2018.04.049

[58] F. Dasí, “Alpha-1 antitrypsin deficiency”, *Med. Clin.* (English Edition), vol.162, no. 7, pp. 336-342, 2024, doi: 10.1016/j.medcle.2023.10.026

[59] M.L. Ivanova, N. Russo and K. Nikolic, “Hierarchical Functional Group Ranking via IUPAC Name Analysis for Drug Discovery: A Case Study on TDP1 Inhibitors”, *ArXiv,* 2025, doi: 10.48550/arXiv.2503.05591

[60] S. Kim et al., “PubChem Substance and Compound databases”, Nucleic Acids Research, 44, D1202-13, 2016, doi: 10.1093/nar/gkv951

[61] M.L. Ivanova, N. Russo and K. Nikolic, ”Predicting novel pharmacological activities of compounds using PubChem IDs and machine learning (CID-SID ML model)”, *ArXiv,* 2025, doi: 10.48550/arXiv.2501.02154

[62] National Center for Biotechnology Information, "PubChem Bioassay Record for AID 1996, Aqueous Solubility from MLSMR Stock Solutions, Source: Burnham Center for Chemical Genomics", *PubChem* [online] https://pubchem.ncbi.nlm.nih.gov/bioassay/1996 (access 20 March 2025)

[63] A. Ortiz-Perez et al., “Machine learning-guided high throughput nanoparticle design”, *Digital Discovery*, vol.3, pp. 1280-1291, 2024, doi: 10.1039/D4DD00104D

[64] L. Yang et al., "Machine learning applications in nanomaterials: Recent advances and future perspectives", *Chem. Eng. J.*, vol. 500, p. 156687, 2024, doi: 10.1016/j.cej.2024.156687

[65] M. Greenacre *et al.,”* Principal component analysis”, *Nat. Rev. Methods Primers, vol.* 2, no.100, 2022, doi: 10.1038/s43586-022-00184-w

[66] GitHub, “Reordering of the feature importance list according to the relative proportion of the active cases” [online] https://github.com/articlesmli/NMR_ML_CHOP/blob/main/8.reordering_feature_importance_list.ipynb (accessed 08 June 2025)

[67] GitHub, “Extraction of the most desirable functional group/fragment for CHOP inhibition” [online] https://github.com/articlesmli/NMR_ML_CHOP/blob/main/9.5.CHOP_group_all_dfs.ipyn (access 08 June 2025)

[68] GitHub, “Extraction of the most desirable functional group/fragment for CHOP inhibition” [online] https://github.com/articlesmli/NMR_ML_CHOP/blob/main/9.5.CHOP_group_all_dfs_ZEROS.ipynb (access 08 June 2025)