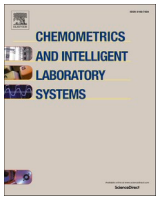
Chemometrics and Intelligent Laboratory Systems 251 (2024) 105168

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Enhancing Hansen Solubility Predictions with Molecular and Graph-Based Approaches

Darja Cvetković *, Marija Mitrović Dankulov, Aleksandar Bogojević, Saša Lazović, Darija Obradović

Institute of Physics Belgrade, National Institute of the Republic of Serbia, Pregrevica 118, 11080 Belgrade, Serbia

A R T I C L E I N F O

Original content: Code and data for XGBoost and GNN prediction of Hansen Solubility Parameters (Original data)

Keywords:
Hansen solubility concept
Graph vs. Molecular solution
Fundamental interpretation
Practical validation
Graph Neural Networks
Machine Learning

A B S T R A C T

The fast and accurate prediction of Hansen solubility benefits many diverse fields such as pharmaceuticals, the food industry, and cosmetics. To estimate the individual HSP values (polar, dispersive, and hydrogen bonding components), we investigated the performance of using Mordred descriptors in multiple linear regressions and XGBoost modeling. For HSP predictions, we also tested a graph-based molecular representation with graph neural network (GNN) modeling. To select the optimal models for final training and predictions, we used nested cross-validation and hyper-parameter optimization. The models with the best predictive performance were selected through internal (R^2_{train}, RMSE, MEPcv) and external (RMSEP, CCC, MEP, R^2_{test}, ar^2m, Δr^2m) validation metrics using ~1200 compounds from free-available database https://www.stevenabbott.co.uk. To confirm the practical reliability, we examined the agreement of experimentally obtained HSP data from the literature for 93 compounds and the data predicted by the created models. The results of GNN modeling showed the best predictive characteristics, which include a coefficient of determination between experimentally obtained and predicted HSP values greater than 0.76 for polar and hydrogen bond forces and greater than 0.66 for dispersive forces. Interpreting the fundamental basis of Hansen solubility using the created MLR equations and XGBoost models, HSP values were found to be influenced by van der Waals volume characteristics, 2D matrix molecular representation, and polarity. We elaborated on the practical benefits of using the selected GNN method through Hansen's solubility sphere as an example. This is the first study to demonstrate the advantages of GNN in predicting individual HSP components, as well as the first study to describe in detail their molecular basis using MLR and XGBoost modeling.

1. Introduction

The solubility of compounds in water is one of the key physical properties that indicates the strength of molecular interaction between solvent molecules, which directly affects a wide range of phenomena important for determining and predicting the fundamental properties of materials and drugs. The fast and accurate solubility prediction benefits many diverse fields such as pharmaceuticals, the food industry, and cosmetics. Commercial chemical products are manufactured as multi-component chemical mixtures, so a basic knowledge of the miscibility of ingredients is required to meet environmental, shelf life, and product quality specifications [1]. Therefore, to efficiently search for satisfying formulations, a predictive tool for solubility is indispensable. The solvent and solute have similar solubility parameters if they tend to be soluble [2]. Several methods that differ in their basic characteristics can be used to predict the solubility of compounds in water including ESOL, LSER, and three-dimensional Hansen solubility evaluation systems. ESOL (ESOL - Estimated SOLubility) is a simple method for estimating the aqueous solubility of a compound directly from its structure against several molecular properties (clogP, molecular weight, rotatable bonds, aromatic proportion, non-carbon proportion, H-bond donor count, H-bond acceptor count, polar surface area) [3]. The LSER is another model developed by Abraham consisting of five solute properties such as excess molar refraction, dipolarity/polarizability, hydrogen bond acidity and basicity, and McGowan volume [4].

A slightly more complex approach defines the total solubility (δ) through a three-dimensional concept, which in addition to the total solubility includes its hydrogen bonding (δ_h), polar (δ_p) and dispersion (δ_d) components [5]. As such, it differs from the previous ones that represent a one-dimensional evaluation system (logarithm of aqueous

* Corresponding author.
E-mail address: darja@ipb.ac.rs (D. Cvetković).

https://doi.org/10.1016/j.chemolab.2024.105168
Received 1 March 2024; Received in revised form 17 June 2024; Accepted 19 June 2024
Available online 21 June 2024

solubility, logS).

The concept of a solubility parameter (δ) was introduced by Hildebrand and Scott, who proposed that solvents with similar δ values would be miscible [6]. The Hildebrand model utilizes a single parameter (δ) defined as the square root of the cohesive energy density, to determine whether a substance is a good solvent or nonsolvent for a selected compound [7]. Hansen proposed an extension of the Hildebrand parameter method to estimate the relative miscibility of polar and hydrogen bonding systems. Therefore, the Hansen solubility model (HSP) uses a three-parameter (3D) concept to quantify the solubility of compounds [8]. In Hansen's approach, the Hildebrand solubility parameter is split into three components: polar, dispersion, and hydrogen bonding forces [5,8].

Using Hansen's solubility parameter values (δ, δ_h, δ_p, δ_d), one can directly evaluate the pharmacokinetic potential of compounds (gastrointestinal absorption characteristics, skin permeation), select appropriate solvents for chemical analysis, and more precisely characterize pharmaceutical formulation, which is not possible only via logS [5].The pharmaceutical analysis can be successfully performed by considering unique physico-chemical and mechanical properties that can offer significant advantages in understanding compound stability, and *in vivo* performances. In pharmaceutical sciences, HSPs are an indispensable physico-chemical concept that has been used to predict the miscibility of a drug with excipients/carriers in solid dispersions [9], as well as for pharmacokinetic studies including prediction of compound skin penetration rate [5] and its oral gastrointestinal absorption [10,11]. A deep understanding and predicting solubility is challenging since it depends on the interaction between solute and solvent, along with various other physico-chemical properties. Different experimental methods were used to estimate the solubility parameter such as: swelling measurements [12]; viscosity measurements [9] and inverse gas chromatography (IGC) [13–15]. Hildebrand and Hansen solubility parameters can be estimated using general computational tools. Choi and Kavasallis first used atomistic simulations to estimate the solubility parameters of a class of alkyl phenol ethoxylates [16] and later applied it to estimate 3D Hansen solubility [17]. A related method has been applied to estimate the HSP forces of fat and oils-related materials [18]. It has previously been found to calculate HSP using molecular dynamics and structure interpolating group contribution (GCM) methods that explore discrete molecular interaction models and require knowledge of the appropriate interaction parameters to calculate thermodynamic properties [19–23]. A method based on lattice fluid theory was developed by Panayiotou [24] in order to estimate the hydrogen bonding component of the Hansen solubility. In ref. [25] a new and direct approach was proposed to estimate the acidic and basic components of the hydrogen bond solubility based on the σ-moments of the screening charge distribution or the sigma profile of the quantum mechanics-based COSMO-RS theory.

In general, novel modeling approaches can differ between those using molecular descriptors for learning (e.g., XGBoost [26]), and graph-based models, (e.g., Graph Neural Networks (GNNs) [27]), that utilize both the structural information from the graph representation of a molecule, along with the information from individual atoms and bonds. Many quantitative structure-property relationships (QSPR) approaches are data-driven and based on machine learning (ML), aiming to find statistical relationships between molecular information and experimental results to produce predictions of important physico-chemical properties [28–32]. For solubility determination, alternative calculation methods are based on combined first principles of quantum mechanics and artificial neural networks (ANN) [33], and predictive modeling based on Gaussian processes and Bayesian ML [34]. Different ML methods were used to estimate the solubility of acid gasses in ionic liquids [35]. [36]In the realm of computational solutions rooted in graph theory, promising outcomes have been achived in various predictive tasks. These include the prediction of ionic conductivity of ionic liquids [36], solvation free energy for solute-solvent pairs [37], aqueous solubility [38], and infinite dilution activity coefficient for solute-solvent pairs [39]. For Hansen solubility parameters prediction, Sanchez-Langeling et al. [34] employed a Bayesian machine learning approach utilizing Gaussian processes, leveraging various input data such as SMILES strings, COSMOtherm simulations, and quantum chemistry calculations. More recently, Pang J. et al. [40] fine-tuned pre-trained natural language processing (NLP) models (Mol2Vec and ChemBERTa) to derive molecule embeddings from SMILES strings, utilizing them for downstream prediction of HSPs. Applying the most optimal theoretical or computational solution to predict certain molecular properties can be a reliable alternative to demanding experimental testing while providing a basic molecular understanding.

In this research, we investigate the efficacy of Genetic Algorithm (GA) for descriptor selection in Multiple Linear Regression model (MLR), XGBoost, and Graph Neural Networks (GNNs) analysis, to estimate 3D Hansen solubility of 1192 model compounds. Apart from the standard prediction metrics, the prediction accuracy of the created models is considered using Roy-metrics to create the strictest criteria for accepting any model in terms of external predictability [41]. In addition, we used experimentally obtained HSP values from the literature of 93 compounds from ref. [5] and ref. [1] and tested the practical ability of the created models by looking at the agreement between the experimentally obtained data from the literature and the predicted data from our models. We based physicochemical interpretation of 3D Hansen solubility on the most successful models that meet the all validation criteria. The practical benefits of the selected models for solvent selection and pharmacokinetic drug optimization were demonstrated using the fundamental characteristics of the 3D Hansen solubility sphere. The created models will serve as a basis for predicting and understanding the HSP concept of solubility.

2. Materials and methods

In this section we provide a theoretical background of the three-dimensional Hansen solubility, describe datasets, models and evaluation metrics. The code, models and the data used in this work are freely available at GitHub https://github.com/darjacvetkovic/HSP-predictions.

2.1. Theory of three-dimensional Hansen solubility

The basic equation governing the assignment of Hansen parameters is that the total cohesion energy, E, is a sum of the individual dispersive (E_d), polar (E_p), and hydrogen bonding (E_h) energies:

$$E = E_d + E_p + E_h \tag{1}$$

Dividing Eq. (1) by the molar volume (V) gives the square of the total (or Hildebrand) solubility parameter (δ (MPa)$^{1/2}$) as the sum of the squares of the Hansen dispersive (δ_d), polar (δ_p) and hydrogen bond (δ_h) component:

$$E / V = E_d/V + E_p/V + E_h/V \tag{2}$$

$$\delta^2 = {\delta_d}^2 + {\delta_p}^2 + {\delta_h}^2 \tag{2a}$$

Quantitatively, the solubility differences between two compounds, compound (1) and compound (2) can be defined by following R_a [MPa$^{1/2}$] value:

$$R_a^2 = 4(\delta_d\,(1) - \delta_d\,(2))^2 + \left(\delta_p(1) - \delta_p(2)\right)^2 + (\delta_h\,(1) - \delta_h(2))^2 \tag{3}$$

Quantitatively, the HSP distance between two compounds, compound (1) and compound (2) can be defined by Ra [MPa1/2] parameter value. Conversely, a large value of Ra indicates that compound (1) and compound (2) will have a different solubility profile.

The Hansen solubility sphere method [5,42,43] is based on the principle that like dissolves like. The interaction radius R_0 [MPa$^{1/2}$]

represents the radius of the sphere of a selected molecule. The ratio between R_a and R_0 has been called the *RED* number reflecting the relative energy distance between two molecules in the Hansen space:

$$RED = R_a/R_0 \quad (4)$$

By definition, RED = 0 is equivalent to no energy difference, RED < 1 indicates high solute-solvent affinity, RED > 1 indicates low affinity and RED = 1 (or around 1) reflects a boundary condition [5]. The compatibility of a solvent with a known HSP categorizes solvents as good or poor.

2.2. Datasets

Here we describe the datasets used for training and validation, and for the confirmation of the practical reliability of our models. Specifically, we use the Hensen dataset for training and validation, while additional datasets from refs. [1,5] are used to further confirm the reliability of MLR, XGBoost and GNN models.

a) The Hansen dataset

To train our models, and to get initial evaluations of their performance we use the 80%:20% train-test split (Table 1) on the dataset of 1192 small molecules from a freely available web database https://www.stevenabbott.co.uk (further referred to as the Hansen dataset).

This dataset consists of 1192 structurally diverse small molecules such as normal alkanes, cycloalkyl compounds, alkenes, alcohols, heterocycles, and their experimentally determined HSP values at 298K. The distributions (histograms) for the number of atoms, and molecular weights of these molecules can be found in Fig.1, while the distributions for Hansen solubility parameters and total solubility are shown in Fig.2.

Molecular descriptors for these compounds, used for XGBoost and MLR models, were calculated using the Mordred descriptor tool for Python [44], while the molecular graphs and their atomic and bond features that were used for GNN models (see subsection 2.3), were generated using the deepchem library for Python [45,46].

Specifically, to get the molecular descriptors for training XGBoost and MLR models, we calculated 1613 2D molecular descriptors using the Mordred descriptor tool. The number of descriptors was further reduced by omitting the descriptors that had zero values for over 90% of the compounds in the dataset. Furthermore, since the calculator can generate various technical errors, the descriptors that had errors for over 5% of the compounds were also discarded. Since the dataset is smaller in size, it is commendable to minimize the number of errors, which is why we used this type of strict condition.The remaining errors are managed by XGBoost's ability to handle missing data and outliers through optimization during training, and the Genetic Algorithm's feature selection process, which focuses on the most informative features, thus mitigating the impact of errors. Additionally, the number of features (the number of molecular descriptors) also exceeds the size of the dataset, so it is generally advised to reduce this number in a systematic way. The outlined process generated 869 descriptors, which underwent additional selection tailored to each model. In the case of MLR models, a genetic algorithm was applied for feature selection utilizing the QSARINS software. For XGBoost models, descriptors yielding optimal performance were chosen within a nested cross-validation loop, alongside hyperparameter optimization (refer to subsection 2.3.1 and 2.3.2 for details).

Table 1
Data split and description table for the Hansen dataset.

Data	Number of molecules	Description
Train	953	Data used for training the models. Training is done in a 5-fold cross-validation loop, where this dataset is further split into the training and validation subset.
Test	239	Data used after training on the Train set to evaluate model performance.

To get the atomic and bond features required for training GNN models, the procedure was more straightforward as no additional feature reduction is required since the featurizer in the deepchem module generates 9 fundamental atomic features and 4 fundamental bond features. These features are encoded into corresponding feature vectors of length 30 and 11, respectively (refer to subsection 2.3.3 for more information). By harnessing these features alongside the "message passing" mechanism inherent to GNNs, the model can effectively integrate information regarding the local neighborhood from each atom and bond within a molecule during training. This integration results in a comprehensive molecular feature vector that encapsulates both topological and chemical properties.

b) Additional datasets

In order to assess the reliability of the created models in practical analysis, the predictions of HSP parameters of all models were compared with the experimentally obtained results from:

1. ref. [1] which contains the experimentally obtained total solubility (δ), and
2. ref. [5] that contains experimentally obtained all HSP values (δ, δ_h, δ_p, δ_d)

Molecules from [1] and [5] that were in the initial set from https://www.stevenabbott.co.uk were excluded from evaluation. Furthermore, for each XGBoost model, the molecules that yielded any error in the calculation of their descriptors were removed from the dataset. This leaves us with 31 molecules (or 27 for XGBoost30 or XGBoost50 models) for testing from [5] and 62 molecules from [1]. The histograms for the number of atoms, molecular weights, Hansen solubility parameters, and total solubility parameters are shown in Fig. 3 , Fig. 4.

2.3. Models and training

The quantitative structure-property relationship (QSPR) methodology was used to define the physico-chemical basis of the 3D Hansen solubility (HSP) concept and to provide models with reliable predictive performances. We use experimentally determined HSP values of a general set of 1192 small molecules from a freely available web database, https://www.stevenabbott.co.uk (the Hansen dataset). Molecular descriptors for these compounds, used for XGBoost and MLR models, were calculated using the Mordred descriptor tool for Python [44], while the molecular graphs and their atomic and bond features that were used for GNN models, were generated using the deepchem library for Python [44,45], as described in section 2.2. We train separate models for each Hansen solubility parameter: δ_d, δ_p, δ_h. This approach enables the models to learn the optimal internal parameters and/or descriptors(in the case of XGBoost and MLR models), particularly tailored to each of these three solubility components. MLR models were developed in the program QSARINS [47,48], while XGBoost and GNN models were developed in Python.

Models are trained and evaluated on the Hansen dataset, using a random 80%:20% train test split (Table 1). We examined the predictive quality of the obtained models by calculating the parameters of internal and external validation. The internal validation was performed on the training set and includes the coefficient of determination R^2_{train} , root mean square error (RMSE), and Mean Error of Prediction (MEPcv). The external validation refers to the test set and includes the coefficient of determination R^2_{test}, root mean square error of prediction (RMSEP), Mean Error of Prediction (MEP), and concordance correlation coefficient (*CCC*). The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. RMSE refers to the training set, and RMSEP refers to the test set.

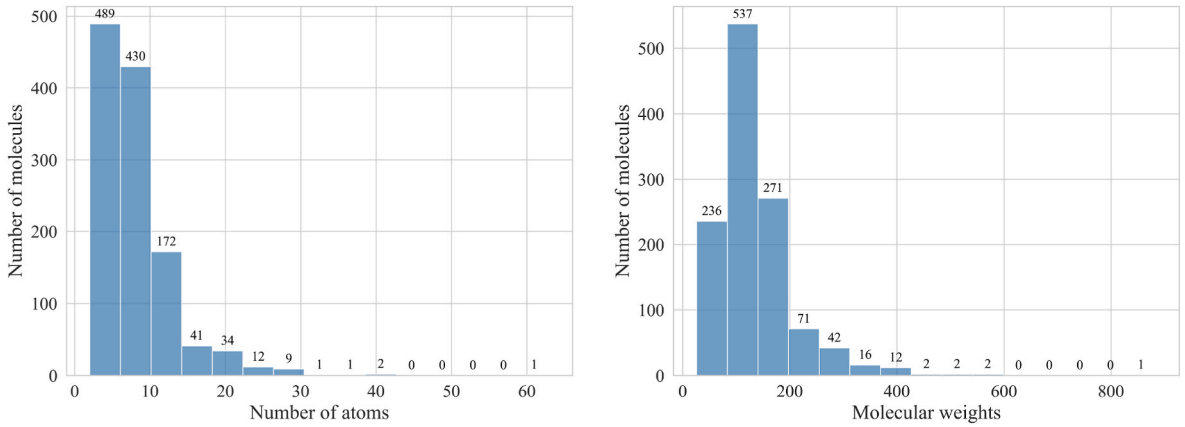


Fig. 1. Histograms of the number of atoms (left) and molecular weights (right) for the training dataset for the Hansen dataset.

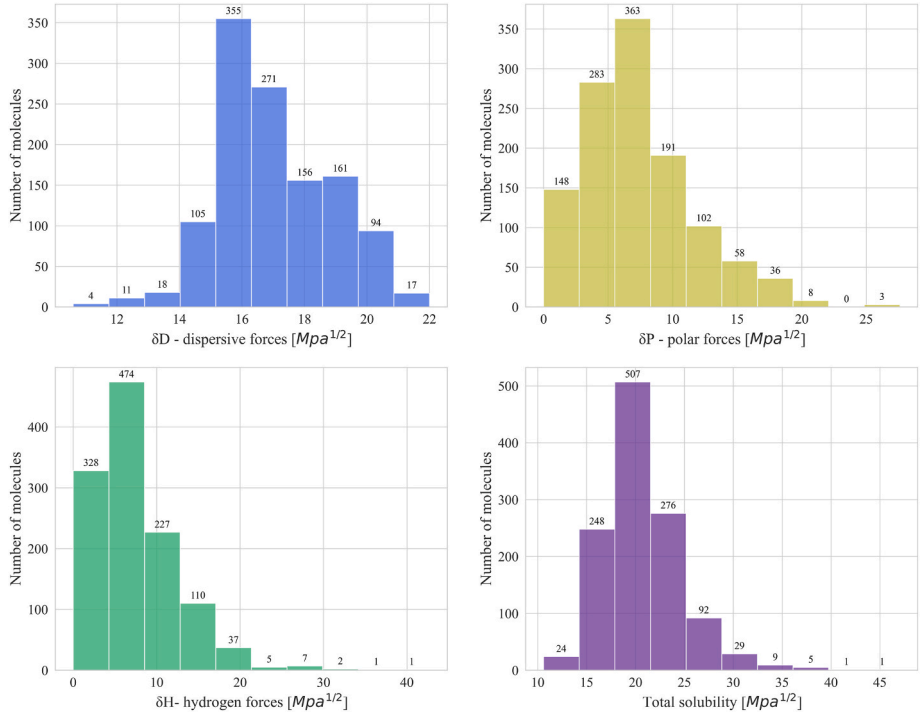


Fig. 2. Histograms for each of the Hansen solubility parameters and total solubility for the Hansen dataset.

A close value of these errors indicates better characteristics of the model. Except for the standard used MEP and R^2_{test}, and *CCC*, the prediction accuracy of the created models is considered by using the parameters of Roy-metrics (ar^2_m, Δr^2_m). This introduces an additional more rigorous way of evaluating the created model. The Roy metrics are significantly different from other external validation metrics, among which *CCC* is the most optimistic. Parameter ar^2_m along with Δr^2_m (lower than 0.2) provides the most stringent criterion of external validation, and as for a minimum number of trials, these criteria are met at a given threshold value. In ref. [41], the use of ar^2_m together with Δr^2_m has been confirmed as the most stringent criterion for accepting any model in terms of external prediction [49].

2.3.1. Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is one of the most commonly used modeling methods in QSPR modeling because they are simpler and easier to interpret. We apply Genetic Algorithms (GA) for descriptor selection which is a very effective procedure, widely and successfully applied in many QSPR approaches [50]. Molecular descriptors were calculated using the Mordred Python module, and the descriptors with intercorrelation over 0.90 were excluded from the analysis.The best-fitted MLR models were selected by the GA in the program QSAR-INS. The GA is a stochastic method to solve optimization problems defined by fitness criteria applying Darwin's evolution hypothesis and different genetic functions, i.e., crossover and mutation [51,52]. The leave-one-out cross-validation (LOO-CV) technique was used to assess the performance of the resulting model. The correlation coefficient of this procedure (Q^2LOO) was calculated and the best model was selected based on the highest Q^2LOO. This coefficient must be close to the value of R^2 to prove that the model is not dependent on the data set [53].

2.3.2. Extreme Gradient Boosting (XGBoost)

XGBoost, or Extreme Gradient Boosting [26], is a machine learning algorithm that uses a type of ensemble learning that combines the predictions of multiple "weak" learners to build a strong predictive model. These weak learners are typically decision trees. Decision trees

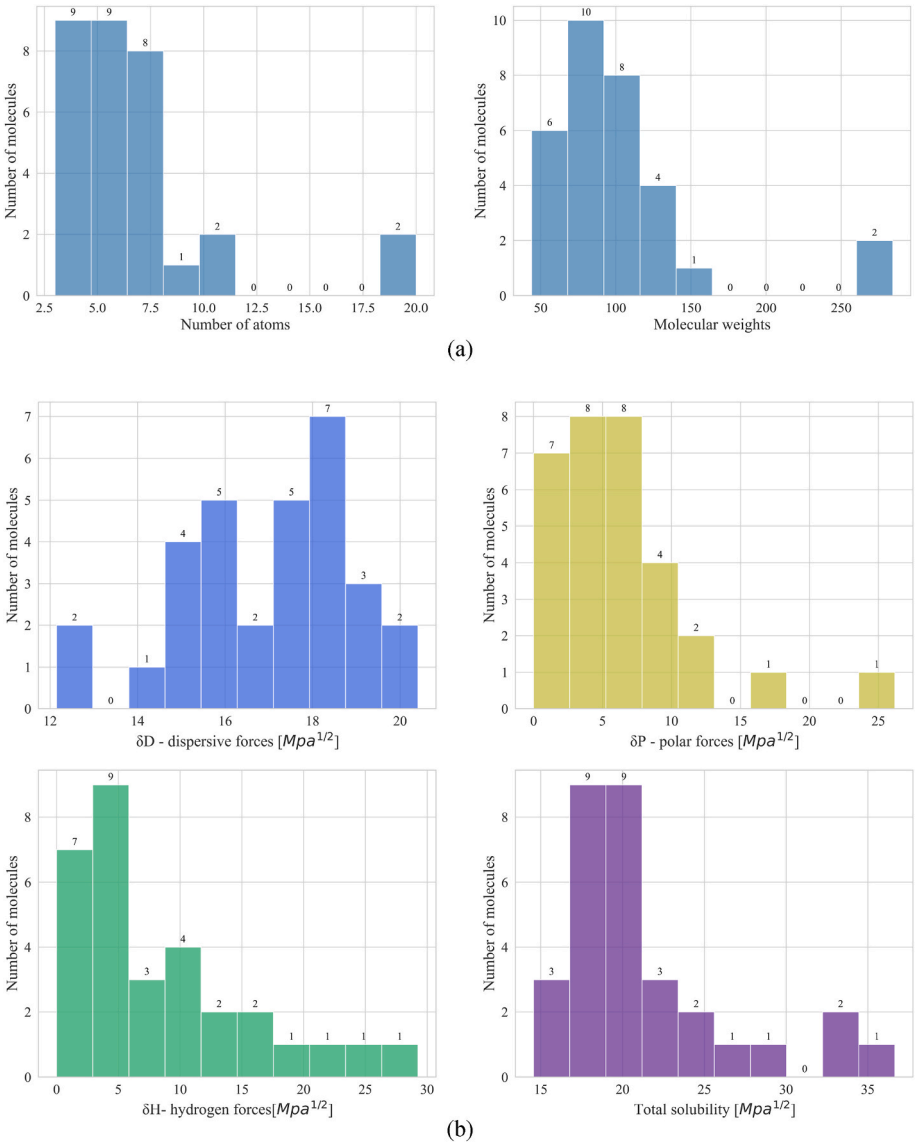


Fig. 3. Histograms of the number of atoms and molecular weights (a) and Hansen solubility parameters (b), along with total solubility for molecules in ref. [5]

systematically split data based on the most informative features (in our case molecular descriptors) and thresholds. In this way, they recursively create a tree-like structure, with each branch leading to a final prediction at a leaf node, ultimately guiding the decision-making process based on the characteristics of the data. XGBoost optimizes model performance by iteratively training trees to correct the errors of previous trees, which effectively boosts their predictive ability. Additionally, the algorithm includes regularization features that prevent overfitting. Due to its predictive power, low computational cost, and ease of implementation XGBoost has gained researchers' and practitioners' attention and was successfully employed in property prediction tasks [54,55].

In this work, we created and compared predictive models with 10 (XGBoost10), 30 (XGBoost30) and 50 (XGBoost50) molecular descriptors (features). The most important descriptors for model predictions were identified using a simpler XGBoost model, and these descriptors were then employed to train more robust models. For each case, a two-level nested cross-validation was implemented, as detailed in the following paragraphs. We do not exclude intercorrelated descriptors prior to training, as some correlated descriptors may represent different properties that, despite their correlation, do not influence each other, so excluding one in favor of another may not be beneficial.

Firstly, in the 5-fold outer loop (which splits the training dataset into a training subset and validation subset for each fold), a "weaker" XGBoost model was initialized and trained for each Hansen solubility parameter, to select the most important features for its solubility prediction (either 10, 30 or 50 features for XGBoost10, XGBoost30 or XGBoost50, respectively). The training of these weaker models incorporated inner cross-validation (3-fold split) with a randomized search on hyperparameters the *max_depth, colsample_bytree, reg_alpha, reg_lambda* parameters. The aim of this search is to find the values of these hyperparameters that yield the best predictive results, adding additional confidence in feature selection. The *max_depth* parameter determines the maximum depth of a tree, with larger values indicating a more complex model (albeit with a higher risk of overfitting), *colsample_bytree* determines the ratio of columns (descriptors) when constructing each tree, while *reg_alpha* and *reg_lambda* are L2 and L1 regularization parameters respectively, with increasing values making the model more conservative (less prone to overfitting). Following this, the most important features (10, 30, or 50) for each Hansen solubility parameter were selected based on the training of these "weaker" XGBoost models. Note that feature importances are automatically calculated for XGBoost models, and can be obtained via the "feature_importances_" member variable on the trained model.

Secondly, after selecting the most important descriptors for each

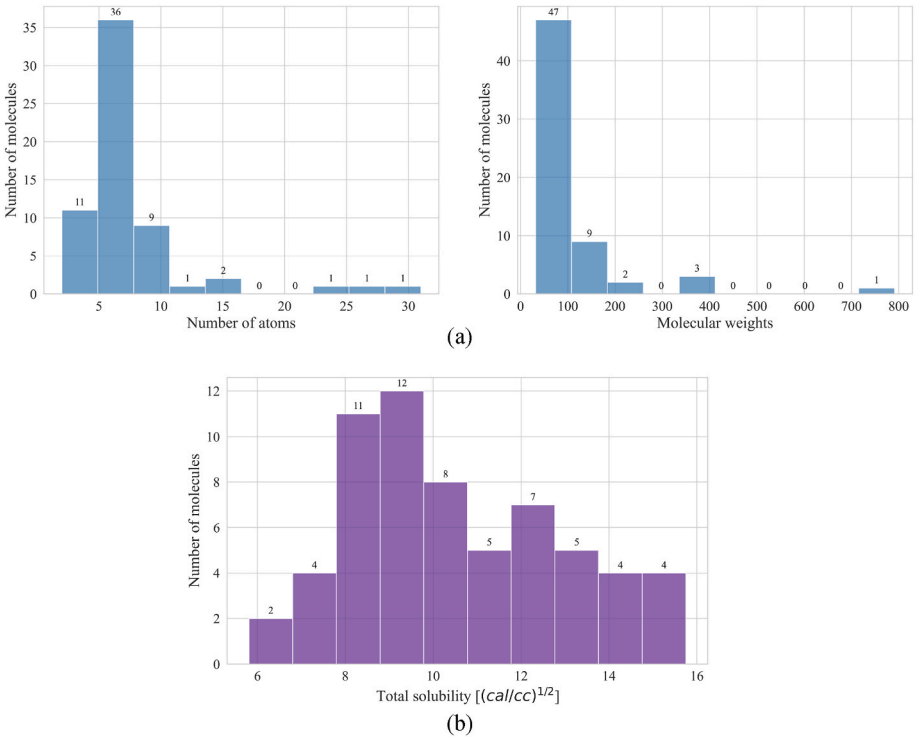


Fig. 4. Histograms of the number of atoms and molecular weights (a) and total solubility (b) for molecules in ref. [1]

parameter, "stronger" models are initialized and trained using the hyperopt Python module [56] for distributed asynchronous hyper-parameter optimization. The parameters that are optimized are *max_depth, gamma, reg_alpha, reg_lambda, colsample_bytree, min_child_weight*. Here, the *gamma* parameter represents minimum loss reduction required to make the next split on a leaf node of the tree, with higher values indicating a more conservative model. Additionally, *min_child_weight* corresponds to the minimum number of instances that need to be in each node - larger values indicating a more conservative model. The performance of each "stronger" model is evaluated on the validation subset from the outer cross-validation loop, and its hyperparameters, R^2 score, RMSE score, and features (descriptors) used for training from the previous step are recorded. Finally, at the end of the 5-fold outer cross-validation loop, for each hyperparameter we are left with 5 models and descriptor sets that yielded the best results inside their corresponding loop iteration (that is, outer cross-validation split). The performance of these models is then evaluated (for each HSP) and models with best performance and the corresponding molecular descriptors used are selected for final training on the whole training set (remember that the outer cross-validation loop split the training dataset into a training subset and validation subset).

In summary, the "weaker" XGBoost models are trained to select the optimal number of molecular descriptors that the model itself considers most important. Both the feature selection and hyperparameter optimization were applied in the cross-validation loop to achieve more confidence, yielding features that were used for training the final predictive XGBoost models.

2.3.3. Graph Neural Networks

Graph Neural Networks (GNNs) [25] are machine learning models specifically designed for learning on graph structures. Graphs are structures that consist of *nodes* and *edges* that connect them, and as such they are used to represent molecules [27]. The molecular graph is the most natural representation of a molecule. Each graph of a molecule can be represented as $G = (\nu,\varepsilon)$, where ν is the set of atoms and ε is the set of bonds. $\nu_i \in \nu$ is the *i*-th atom and $e_{ij} \in \varepsilon$ is the chemical bond between the *i*-th atom and the *j*-th atom [34]. A graph-based representation of the molecule and the graph-learning mechanism are given in Fig. 5. In our research, we use the simplified molecular-input line-entry system (SMILES) [48] strings of molecules to convert molecular structure to molecular graphs. The SMILES of each molecule were converted into a molecular graph using the RDKit toolkit [49] for Python.

Unlike XGBoost and MLR models that utilize molecular descriptors, GNN models use the features of nodes (atoms) and edges (bonds) for learning. They do the learning via the "message passing" mechanism which effectively combines these features for each atom/bond and their local environment [27,29,30]. Depending on the mathematical or methodological formulations for the message-passing mechanism, we can distinguish different GNN models. In this work we train and test four different GNN models for predicting each of the Hansen solubility parameters: the graph convolutional network model (GCN) [57], graph attention network (GAT) model [58], the AttentiveFP model [59] and the message passing neural network (MPNN) model from [60]. Similarly to XGBoost modeling, we perform cross-validation along with hyperparameter optimization and select the model with the best mean performance across all cross-validation folds for final extensive training.The procedure is as follows.

For GNN modeling, we initialize the atomic and bond features [39], and GCN, GAT, AttentiveFP, and MPNN models using the deepchem Python module. Similarly to XGBoost modeling, the four models are trained in a 5-fold cross-validation loop to assess their mean performance on different train-validation data splits. In this process, hyperopt is employed to tune various model hyperparameters for different deepchem PyTorch models. There is no need for an inner cross-validation loop, or "weaker" models since we do not need to further reduce the number of features like in XGBoost modeling. For the GCN model, hyperparameters such as *layer_sizes, dropouts,* and the *residual* parameter are optimized. These parameters control aspects like the size of each layer, neural network node dropout rates, and the inclusion of residual connections, which influence the model's depth and regularization. Similarly, for the GAT model, optimization is done on parameters like *layer_sizes* and *dropouts*, determining the layer sizes and

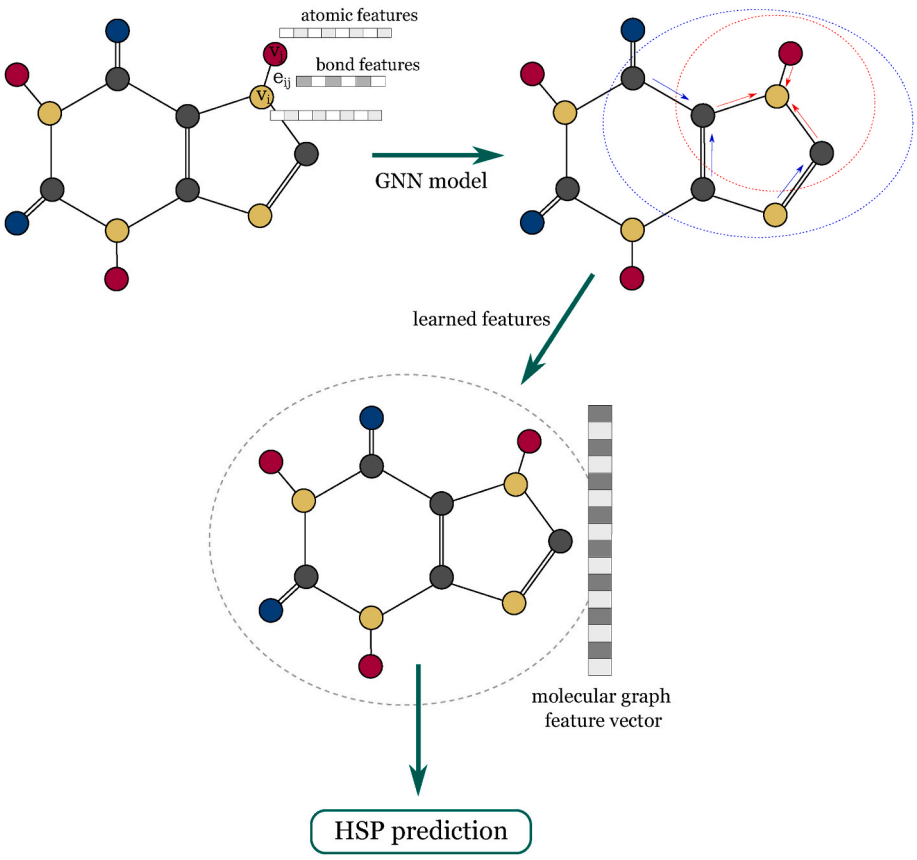


Fig. 5. Graph representation of the molecule and the graph learning mechanism. Each atom is represented by a node and each bond with an edge - their feature vectors are used by the GNN model for learning the final graph feature vector that is used for HSP prediction. Atomic and bond features are combined and updated for each atom/bond with the features from their local environment (this constitutes "message passing" mechanism). The (molecular) graph feature vector is calculated by aggregating individual atomic/bond features which is used to predict HSP parameters.

dropout rates for the network. The AttentiveFP model's hyperparameters include d*ropouts, num_layers, graph_feat_size,* and *num_timesteps*, affecting features such as dropout rates, the number of layers, the size for graph representations and the number time steps for updating the graph representations. Lastly, the MPNN model's hyperparameters involve *node_out_feats*, *edge_hidden_feats*, *num_step_message_passing*, *dropouts*, and *self_loop*, defining node and edge output dimensions, the number of message passing rounds, dropout rates, and the inclusion of self-loops in the message passing process, respectively. GNN training with hyperparameter optimization is time-consuming, which is why, for the purpose of performance assessment with cross-validation, the models were trained for a smaller number of epochs (200) along with early stopping. When the cross-validation loop ends, the mean performance metrics of all four models are compared. The best-performing model for each Hansen solubility parameter was selected for final training with a more extensive hyperparameter optimization search to provide the best results.

3. Results and Discussion

In this section, we present and discuss the results of our study, and compare the performance of our models with those found in the literature. Additionally, we provide a physico-chemical interpretation of the Hansen solubility parameters by analyzing the importance of molecular descriptors used in the XGBoost10 and MLR models. Finally, we demonstrate the practical applicability of our approach through the concept of the HSP sphere, with the predictions derived from GNN models.

3.1. Molecular and graph theory for HSPs prediction

The statistical performance of three proposed methods including XGBoost, MLR and GNNs was tested by predicting Hansen solubility of 1192 structurally diverse set of small molecules from the Hansen dataset (Section 2.2). The most important statistical results are summarized in Table 2. More detailed results containing all evaluation metrics and descriptors can be found in the Supplemenatry Information table.

All created models meet the internal and external validation criteria. GNN models gave significantly better statistical results in the case of the prediction of polar forces than the rest of the models (R^2_{test}= 0.82, ar^2_m = 0.76, and Δr^2_m = 0.06). MLR model showed better predictive performances for dispersive forces (R^2_{test} = 0.85, ar^2_m = 0.78, and Δr^2_m = 0.07), closely followed by the GAT GNN model (R^2_{test}= 0.85, ar^2_m = 0.76, and Δr^2_m = 0.14), while XGBoost models were more suitable for predicting the hydrogen bond forces ($R^2_{test} \geq 0.75$, $ar^2_m \geq 0.66$ and $\Delta r^2_m \leq 0.09$), specifically the XGBoost50 model performed best with R^2_{test} = 0.80, ar^2_m = 0.74 and Δr^2_m = 0.06. Additionally, we observe that XGBoost10, XGBoost30 and XGBoost50 models have similar predictive performances within the respective HSP components (i.e., R^2_{test}, ar^2_m, and Δr^2_m values) which suggests that even 10 molecular descriptors (XGBoost10) per Hansen solubility parameter are enough for successful prediction on this dataset.

We then proceed to test our models on additional datasets from [5] and [1], previously described in the Dataset section. For GNN models we use the AttentiveFP architectures to predict the dispersive (δ_d) and hydrogen bond (δ_h) components, but use the GAT architecture to predict the polar component (δ_p), since those specific architectures yielded the best results for the corresponding components (as presented in Table 2).

Table 2
The most important statistical results of the created models.The external validation is performed on Hansen's dataset test set that includes the following parameters: R^2test, RMSEP, CCC, MEP, ar^2m, and Δr^2m values. The internal validation is based on a training set including R^2, r^2_{adj}, RMSE, and MEPcv values.

HSP	Model	R^2train	RMSE	CCC	MEP	R^2test	RMSEP	ar^2m	Δr^2m
Dispersive forces	MLR	0.85	0.71	0.92	0.53	**0.85**	0.72	**0.78**	0.07
	XGBoost 10	0.87	0.65	0.88	0.65	0.78	0.89	0.70	0.12
	XGBoost 30	0.91	0.54	0.88	0.60	0.78	0.90	0.69	0.10
	XGBoost 50	0.93	0.47	0.90	0.58	0.81	0.84	0.73	0.10
	GNN (AttentiveFP)	0.93	0.43	0.91	0.53	**0.85**	0.80	**0.76**	0.14
Polar forces	MLR	0.60	2.76	0.70	2.37	0.50	3.25	0.50	0.10
	XGBoost 10	0.76	2.09	0.78	1.90	0.65	2.66	0.53	0.22
	XGBoost 30	0.77	2.07	0.79	1.95	0.65	2.64	0.54	0.21
	XGBoost 50	0.88	1.48	0.79	2.00	0.64	2.69	0.52	0.18
	GNN (GAT)	0.89	1.30	0.90	1.31	**0.82**	2.10	**0.76**	0.06
Hydrogen bond forces	MLR	0.73	2.88	0.80	2.12	0.64	3.11	0.53	0.03
	XGBoost 10	0.88	1.80	0.86	0.86	0.75	2.51	0.66	0.09
	XGBoost 30	0.92	1.43	0.88	1.58	0.77	2.40	0.70	0.03
	XGBoost 50	0.90	1.67	0.90	1.43	**0.80**	2.22	**0.74**	0.06
	GNN (AttentiveFP)	0.91	1.44	0.82	1.86	0.67	2.64	0.57	0.07

Additionally, we note that we calculate the total solubility δ from the equation (2a), after getting model predictions for its components. The obtained prediction results are presented in Table 3.

The most successful agreement of the predicted solubility data with the experimental values from the literature was shown by the created GNN models ($R^2 > 0.60$). In the first step, the best correlation with δ [1] was found for the GNN models ($R^2 = 0.76$), which is higher than the correlation obtained by the proposed calculation method in ref. [1] ($R^2 = 0.70$), Fig. 6. The best correlations with experimental partial (δ_d, δ_p, δ_h), and total solubility (δ) parameters of 59 model compounds [5] were also found for the GNN predictions. This confirms the practical validation of the created GNN method, Fig. 7.

Furthermore, we note that the GNN models seem to be the only models that meet the criteria $R^2 > 0.60$, for all predictions, while other models sometimes fail to do so. Specifically, all models other than the GNN model fail to meet this criteria when predicting the dispersive component δ_d from ref. [5], and MLR fails for δ_p [5], while XGBoost10 fails to meet the criteria for δ [1]. The reason for this discrepancy may be that the MLR and XGBoost models overfit the training data which hinders their generalization abilities, but upon closer inspection it appears that this decrease in performance seems to be the case only for the dispersive component δ_d [5] - for the polar and hydrogen bond component in [5] the models seem to have even better performance than for the same components from Table 2 (i.e. the Hansen dataset test set). This is unexpected, and the reason for this may lie in the nature of the dispersive component. It is known that the strength of the London dispersion forces also depends significantly on the shape of the molecule because the shape determines how much one molecule can interact with neighboring molecules at any time, therefore a 3D molecular description could give better results. Even for the GNN model, the decrease in predictions of the dispersive solubility parameter is apparent. Additionally, the decreased prediction performance could be due to the distribution of the training dataset (Figure 2), which contains significantly fewer data points (molecules) with dispersive component values around 12 $MPa^{1/2}$ and 20 $MPa^{1/2}$. As a result, the models may struggle to generalize predictions for such molecules. It is particularly visible that octanoic acid, and glycerol are the prediction confidence outliers (Fig. 7), the removal of which increases the coefficient of determination (R^2) from 0.66 to 0.73. Despite this, the coefficient of determination for the GNN model is still greater than 0.50.

Table 3
The coefficient of determination (R^2) between predicted and experimentally obtained solubilities.

R^2 HSP	MLR	XGBoost10	XGBoost30	XGBoost50	GNN
δ [1]	0.67	0.58	0.61	0.69	**0.76**
δd [5]	0.42	0.45	0.4	0.38	**0.66**
δp [5]	0.56	0.77	0.71	0.85	**0.98**
δh [5]	0.86	0.9	0.94	0.9	**0.96**
δ [5]	0.77	0.76	0.79	0.77	**0.94**

Additionally, we calculate the relative errors of total solubility predictions for each of these test datasets, and for each model - GNN, XGBoost with 10, 30, 50 features and for the MLR model. We present the distributions of these errors in the form of box-plots in Fig 8. Box plots provide a simple overview of the underlying distribution - they display the median, quartiles, and the outliers in the data, which is in our case the relative error in total solubility prediction for each molecule. From Fig 8 (a). We confirm that the GNN model has the best performance on the data from [5] - the median of the relative error distribution is the lowest, while only one of the outliers (molecules) has a relative error more than 20%. On the other hand, from Fig 8 (b). for the data in [1] all the models seem to exhibit very similar performance when it comes to the distributions of relative errors.

3.1.1. Comparison with the results from the literature

Sanchez-Langeling et al. [34] investigated the applicability of Gaussian processes, a Bayesian machine learning approach (gpHSP) for HSP prediction emphasizing the importance of molecular similarity, sigma profile similarity, electrostatic similarity, and shape/size similarity. Multiple models for solubility prediction were compared with reported experimental data where the Pearson's correlation coefficient was obtained within the following range $0.91 > r > 0.37$. Given the disparity in the molecular types utilized in their study compared to ours (small molecules up to 200 atoms, and polymers), direct comparison of these models is challenging. However, it's noteworthy that while the features employed in [34] may encode richer physico-chemical information about the molecules (such as their electrostatic properties), they could entail computational intensity in their calculation - which is not the case for Mordred descriptors. One might consider these differences when deciding which approach to implement. Our results show that XGBoost10, MLR, and GNN modeling meet reliable applicability in practice with a high degree of agreement between predicted values and experimental data from the literature.

More recently, Pang J. et al. [40] utilized Natural Language Processing (NLP)-inspired molecular embedding models to predict Hansen solubility parameters on the Hansen dataset. Specifically, they fine tuned pre-trained Mol2Vec [61] and ChemBERTa models [62,63] to get a feature vector of a molecule from its SMILES representation, and then used those features to predict HSPs. Furthermore, the authors also tested the predictions using Morgan fingerprints as descriptors with XGBoost models for comparison, but the ChemBERTa models have yielded the best results overall (Zinc-base and 77M-MTR). Nevertheless, it seems

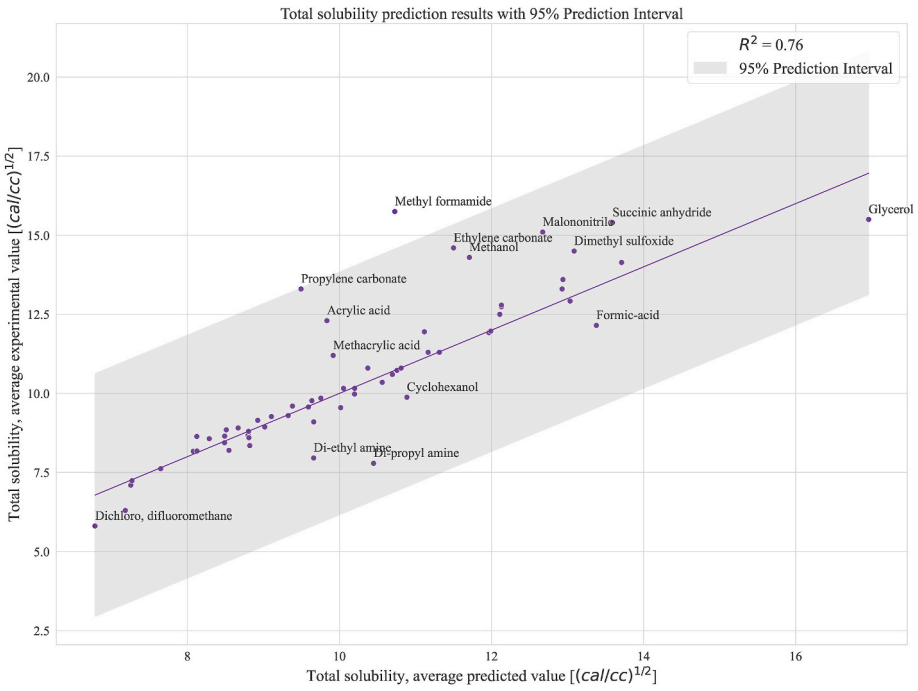


Fig. 6. The correlation between experimental and GNN-predicted δ for the dataset in ref. [1]. The names of the 15 molecules with the highest difference in predictions are shown in the graph. The prediction interval estimates the range within which a future observation is expected to fall with the 95% confidence level.

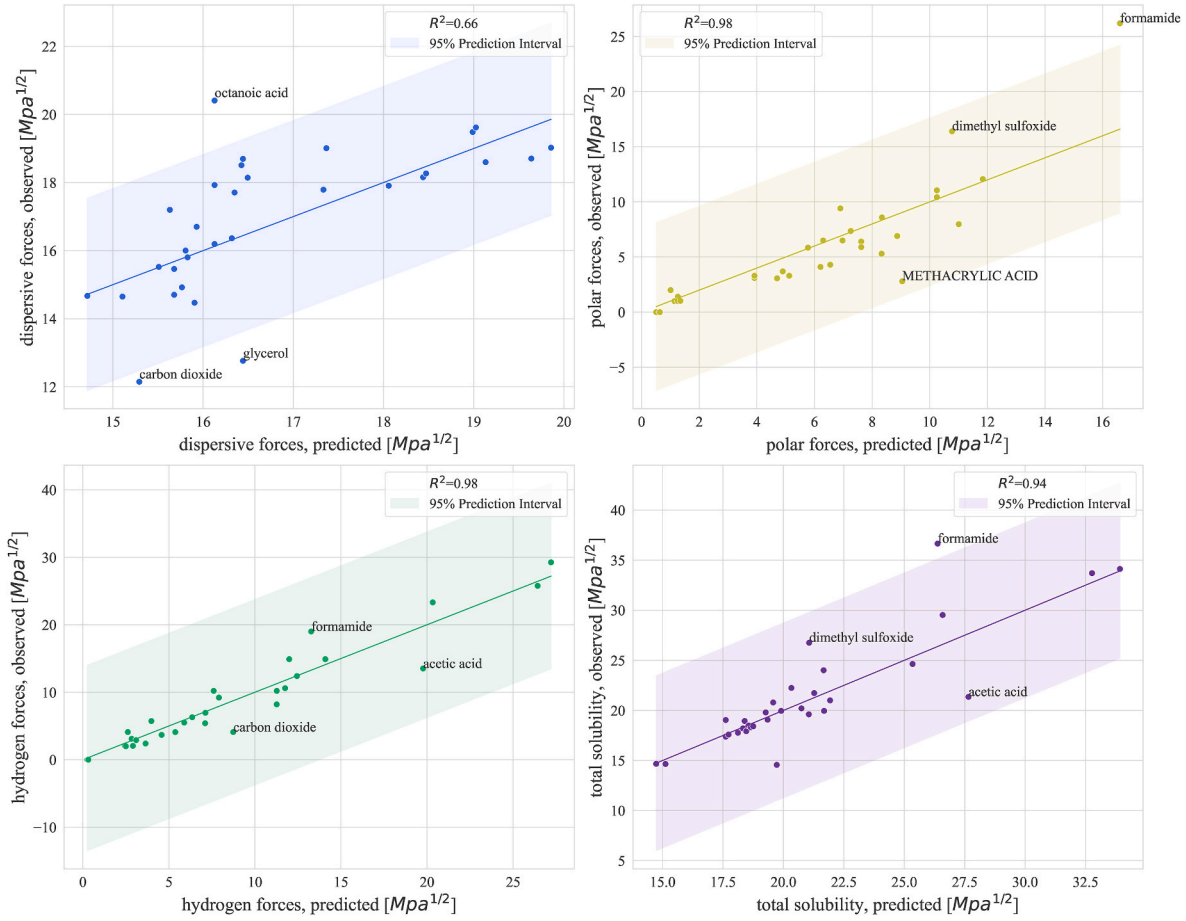


Fig. 7. The correlation between experimental and GNN-predicted HSPs (δ_d, δ_p, δ_h)for the compounds in ref. [5], the 3 molecules with the highest difference in predictions for each solubility parameter are shown in the graphs.

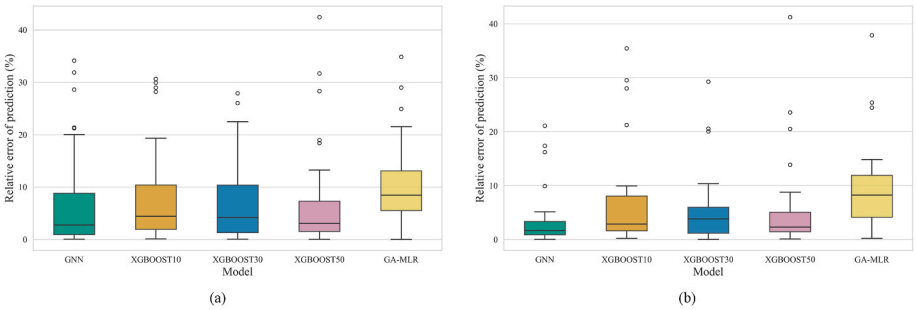


Fig. 8. Box plots of relative errors of predictions of total solubility for each model. (a) Relative errors of prediction on the data from [5] , (b) relative errors of prediction on the data from [1]

that our models perform better on the same dataset (Table 4). In Table 4, we compare the performance of our best models from Table 2, with the best models from [40]. Our models achieve better results than the NLP based models for all Hansen solubility parameters on the Hansen dataset, especially for the polar component (R^2test (GAT) = 0.82 vs. R^2test (Zinc-base) = 0.41). Pang J. et al. suggest that the reason behind the poor performance on the polar component might be due to intrinsic errors in its definition, which limits the prediction accuracy of ML models. However, our results indicate that this may not be the case, as we obtained good prediction results with the GAT model on both the Hansen dataset and the datasets from [1] and [5]. This seems to be the case only for the GAT model, as we observe poorer performance of XGBoost and MLR models (R^2test $\leq$ 0.65). On the other hand, if we compare the performances on all the models from Table 2, the NLP-based model performed slightly better than our GNN (AttentiveFP) and MLR models for the hydrogen bond component on the Hansen dataset (R^2test (77M-MTR) = 0.7, vs. R^2test (AttentiveFP) = 0.67, R^2test (MLR) = 0.64). Still, we have tested our models on two additional datasets [1] and [5], which showed that our GNN models have good generalization capabilities, as well as some of the other models (refer to the previous section for details).

3.2. Physico-chemical interpretation

Through the physico-chemical interpretation based on selected 2D Mordred descriptors, we described the fundamental concept of HSP solubility. In MLR and XGboost modeling, the most significant molecular characteristics that influence the individual HSP values (δ_d, δ_p, δ_h) are discussed separately.

Interpreting GNN models is challenging due to their reliance on basic atom and bond properties, and while methods like identifying important subgraph structures or using interaction map layers for solute-solvent interactions exist, they are outside the scope of this paper but offer promising potential for future research [73-75,37]. Nevertheless, molecular representation and learning based on GNNs have shown excellent accuracy and precision in various fields [31,76–78]. These results show their importance for further advancement of graph based methods for numerical prediction of different physico-chemical properties, and show the need for further development of methods and approaches that would enable their interpretability.

Table 4
Comparison of our best models with the best models from [40]. For comparison with other models refer to Table 2.

HSP	Model	RMSEP	R^2test
Dispersive forces	MLR	0.72	0.85
	GNN (AttentiveFP)	0.8	0.85
	Zinc-base	0.83	0.73
Polar forces	GNN (GAT)	2.1	0.82
	Zinc-base	2.83	0.41
Hydrogen bond forces	XGBoost 50	2.22	0.8
	77M-MTR	2.7	0.7

3.2.1. SHAP feature importance information in XGBoost

In the obtained XGBoost10 modeling, HSP forces are determined by the interplay between geometry, structural and physico-chemical properties of compounds. We use the SHAP values [64] to get the feature (molecular descriptor) importance information from the created XGBoost10 model and understand the molecular basis of 3D Hansen solubility, Fig. 9. SHAP values are based on the Shapley values from game theory, specifically it is a way of determining the contribution of each player in a collaborative game. The players in the machine learning context are features, that is, molecular descriptors for our case. The bar plots in Figure 9 (left panels) show the mean average impact of features for each model (that is, the model for each of the solubility parameters), while the swarm plots, Fig. 9 (right panels), represent more detailed distributions. Each circle in the swarm plot corresponds to a molecule and is colored based on a specific feature's value (ranging from low to high). Red circle positioned on the right side of the scale means that a high feature value positively influences the model prediction, leading to a higher HSP parameter value. Conversely, the blue circle situated on the left side of the scale implies that the feature had a negative impact on the model prediction, resulting in a lower predicted HSP parameter value. Notably, when blue circles (representing molecules with lower feature values) appear on the right side of the scale, it signifies that these lower features had a positive influence on the prediction, and vice versa.

The influence of molecular features on HSP values is discussed in relation to the common property, which can be divided into several classes including features associated with the van der Waals surface, 2D matrix molecular description and polarity. The influence of these properties is discussed below for each individual HSP component (δ_d, δ_p, δ_h).

It is known that ionization can affect the solubility of compounds in certain solvents. In the case of a covalent compound that dissolves in polar solvents (eg, water), no ionization should occur because the force driving the dissolution process would then be van der Waals attraction. Van der Waals attractions were found to be influential in modeling HSPs. The features associated with VSA (van der Waals surface area) represent a volume surface area that can be calculated either for the entire molecule or for parts of the molecule with certain attributes. In the PEOE_VSA parameter, PEOE denotes Partial Equalization of Orbital Electronegativities, which is a charge calculation method, and VSA signifies Van der Waals Surface Area [65,66]. PEOE_VSA6 corresponds to the division of the surface of the molecule conditioned by the atomic partial charge less than -0.05 which was found for polar HSP attractions (δ_p). The descriptor VSA_EState8 is found for the component δ_d indicating the level of accessibility of the atom in the molecule for dispersive forces. VSA_EState8 is equal to the sum of the electro-topological state (E-state) values of atoms with a van der Waals surface area between 6.45 and 7.00 [67]. GATS1v represents the Geary auto-correlation coefficient of van der Waals volume separated by one bond. The SMR_VSA5 descriptor (van der Waals atomic surface where molar

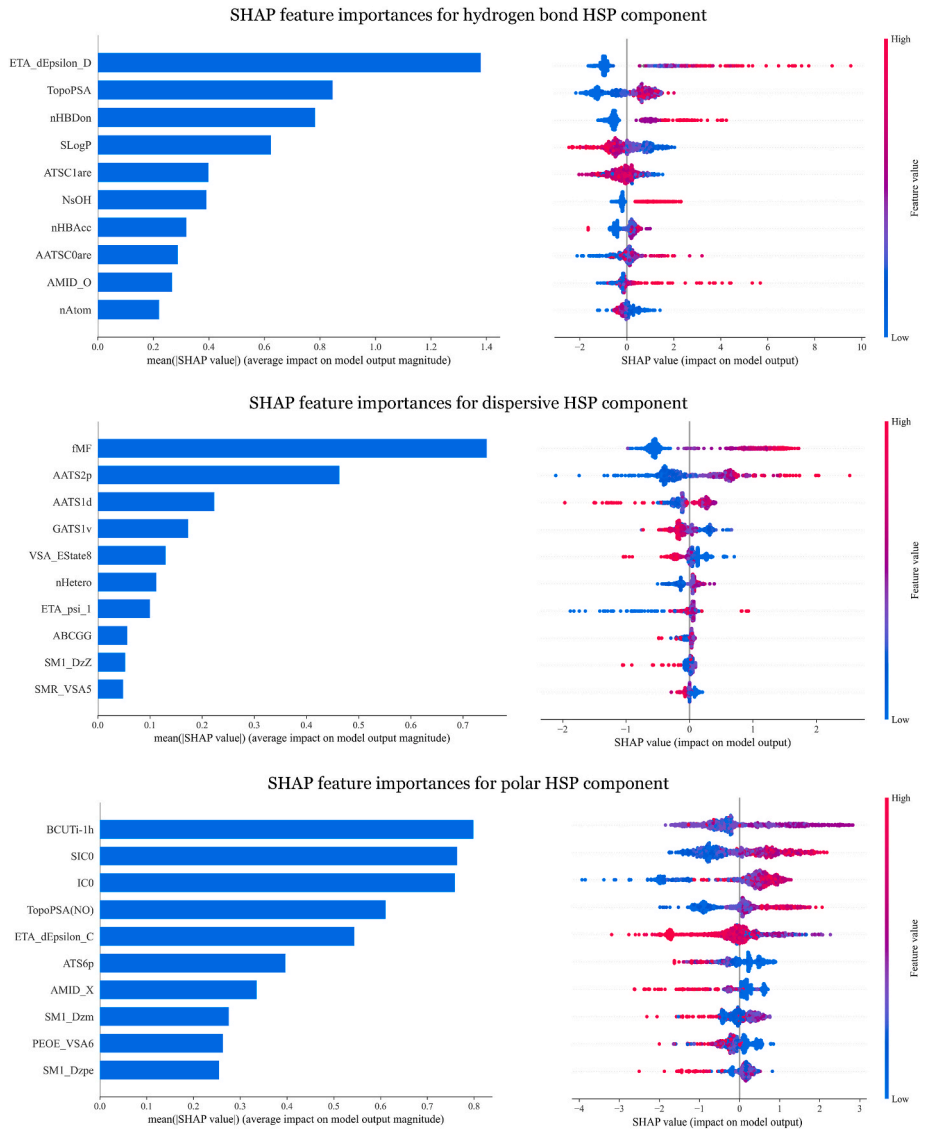


Fig. 9. SHAP feature importances for XGBoost10 models.

refractivity is between 2.45 and 2.75) is the contributor to the dispersive HSP component, which measures the steric factors and bulkiness of the given molecules. SMR_VSA5 is based on molar refractivity, which is directly related to the polarizability of tested molecules and dispersion force strength (δ_d).

SM1_Dz(Z) belongs to a set of descriptors calculated from 2D matrices derived from the molecular graph (2D matrix-based descriptors). This descriptor is the sum of the eigenvalues of the Barysz matrix, whose elements take into account information on both the bond order and the atomic number. In δ_d modeling, this descriptor refers to heteroatoms that influence dispersive HSP forces. The molecules with the lowest SM1_Dz(Z) values are entirely constituted by carbon atoms (both aromatic and not) while the largest values are taken on by highly fluorinated and chlorinated compounds and, more in general, compounds with several heteroatoms. In addition, the number of hetero atoms (nHetero) was also found to have a significant effect on dispersive forces (δ_d). It is known that the greater the difference in electronegativity, the more polarized the electron distribution and the greater the partial charges of the atoms. SM1_Dzpe is the spectral moment of order 1 from the Barysz matrix weighted by Pauling electronegativity, while the SM1_Dzm is weighted by mass. These two descriptors are selected within polar HSP (δ_p) forces. ATS6p is the Broto-Moreau autocorrelation descriptor of lag 6 that is weighted by polarizability. It was found as an important factor for δ_p modeling. BCUTi-1h is the first largest ionization potential-weighted eigenvalue of the loading matrix that was identified as an influential feature for the polar HSP forces (δ_p). BCUT metrics are an extension of Burden's parameters that are based on a combination of an atomic number for each atom and a description of the nominal bond type for adjacent and non-adjacent atoms. These descriptors reflect the influence of bonding information and atomic properties (eg, atomic charge, polarizability, hydrogen bonding abilities relevant to intermolecular interactions) on solubility properties.

The descriptors associated with the polarity of molecules are related to the hydrogen bond and surface area. The contribution of hydrogen bond donor atoms (ETA_dEpsilon_D), the number of hydrogen bond donors (nHBDon), and the number of hydrogen bond acceptors (nHBAcc) are directly associated with the hydrogen bonding forces and are included in polar (δ_p) and hydrogen bonding (δ_h) HSP components. ETA_dEpsilon accounts for the electronegativity contribution of hydrogen bond donor atoms in relation to the electronegativity of heavy atoms in a compound. The nHBDon is calculated as the number of hydrogen bond donors in a compound, which are any −OH or −NH groups where the formal charge of the oxygen or nitrogen is non-negative (i.e., formal charge≥0), while the nHBAcc is the number of

hydrogen bond acceptors in a compound. ETA_psi_1 descriptor appears only once for the model describing the δ_d HSP. This extended topochemical atom descriptor is a measure of the hydrogen bonding propensity of the molecules and/or polar surface area. A higher ETA_Psi_1 descriptor value describes compounds with less electronegative atoms, which are parts of hydrogen bond forming and polar surface area. More polarizable molecules and molecules with larger surfaces have the formation of greater dispersive forces [68]. IC0 descriptor is associated with the size and complexity of molecules, a zeroth order information content index. IC0 is calculated from Shannon's entropy as $-\sum i\ p_i \log_2 p_i$, where p_i is the probability of randomly selecting an atom of a specific type i in the molecule. This descriptor characterizes molecular complexity as the average amount of information per atom type. IC0 is included only for the polar HSP component, which indicates that this descriptor is significant in describing solubility properties for uncharged polar compounds. The positive sign of its regression coefficient from the δ_p/MLR model indicates that compounds containing more different types of atoms (ie, more complex molecules) are more soluble and form stronger polar interactions. For the hydrogen bonding HSP component, topological polar surface area (TopoPSA) becomes less important than hydrogen bond properties (nHBDon, nHBAcc). TopoPSA calculated based on the fragment contribution method, describes the polar surface of the compound, while TopoPSA(NO) uses only nitrogen and oxygen atoms and reverts to the polar HSP component. Compounds with a larger polar surface may be involved in stronger interactions with the polar solvents [69]. The molecular framework fraction (fMF) has the greatest influence on the dispersion HSP forces (δ_d). The descriptor fMF is defined as the number of heavy atoms in the molecular framework MF divided by the total number of heavy atoms (HEVtotal). From δ_d/MLR the solubility increases with increasing fMF and this trend is independent of the ionization state of the molecule. The atom-bond connectivity (ABC) index is one of the recently most investigated degree-based molecular structure descriptors. The relationship between the Graovac–-Ghorbani index of atom–bond connectivity (ABCGC) and the dispersive HSP forces was established in δ_d/XGBoost modeling. A topological descriptor, (SIC01) describes a structural information content of 0-order. It reflects the size and compactness of the molecule. The descriptor SIC01 is based on Shannon information theory and like other molecular complexity indices makes the division of atoms into different classes depending on the size of the coordination sphere taken into account near a given atom. Its effect on solubility including a negative effect on dispersive HSP in MLR indicates that the solubility of the tested compounds decreases with increasing their molecular size. AMID_X is the most influential descriptor in δ_p/XGBoost modeling. AMID_X calculates the averaged molecular weighted path (ID) on halogen atoms. Relatively high electronegativity of halogens gives highly polar interactions of covalent bonds, so a larger number of halogens favors polar HSP interactions. AMID_O was found important for δ_h/XGboost modeling and it represents averaged molecular ID on O atoms which emphasizes the importance of oxygen atoms in the creation of hydrogen bonds. Another significant descriptor that has an impact on the hydrogen-bonding HSP component in XGBoost modeling, NsOH, indicates the number of sOH (–OH) type atoms. This descriptor is related to hydrophilicity. The ATSC1are is based on Allred–Rochow electronegativity between adjacent atoms, and it was selected within hydrogen bond HSP forces (δ_h). Consequently, it could identify polar bonds in the cations and act as a measure of dipolar interaction strength to δ_h forces. It was found that the ratio between molecular hydrophilicity and hydrophobicity (SLogP), as well as the number of all atoms (nAtom descriptor), are more important in predicting solubility (δ_h/XGBoost).

3.2.2. Physico-chemical interpretation from MLR

A hybrid approach combining GA with multiple linear regressions (MLR) was used to define the most suitable predictive MLR dependences to characterize the dispersion (δ_d), polar (δ_p), and hydrogen bonding (δ_h) HSP components. The following MLR predictive models are obtained (Eqs. (5a)-(5c)):

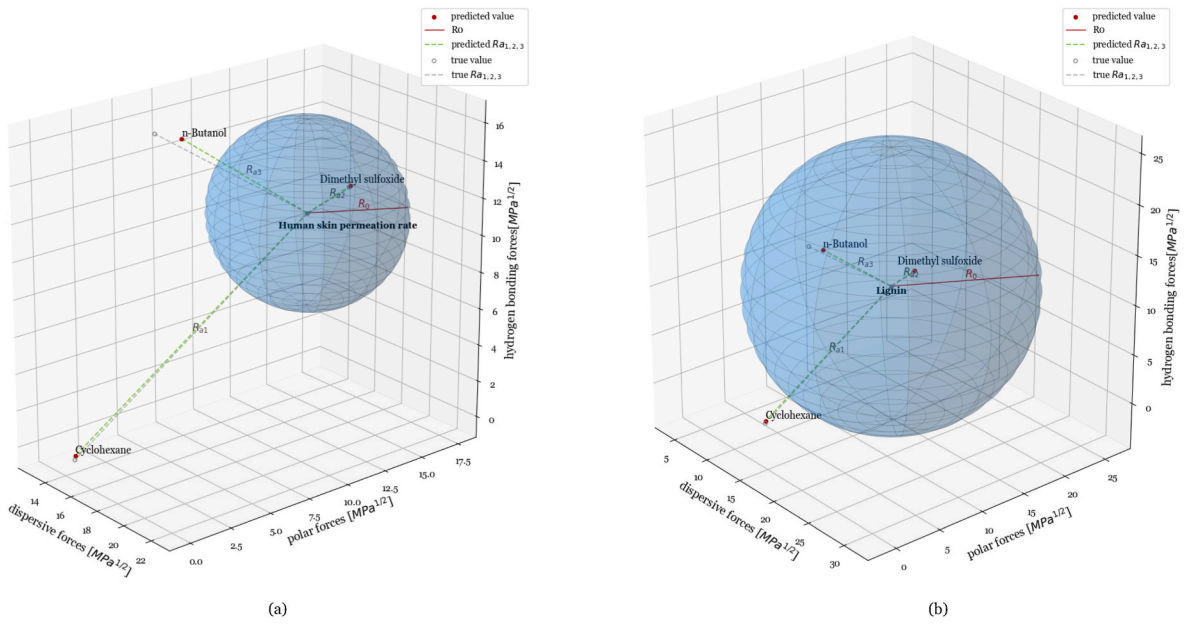
1. $\delta_d = 18.23 - 0.27\ nO - 0.002\ ATS5dv - 0.07\ AATS0i + 0.05\ AATS1i - 0.003\ ATSC3dv + 0.0004\ ATSC2v - 0.03\ ATSC3p + 0.04\ AATSC1v + 0.21\ nBondsKD + 0.19\ C1SP3 - 0.35\ Xc\text{-}3dv + 1.21\ AXp\text{-}1dv + 1.09\ Mm - 0.08\ SdssC - 0.01\ fragCpx + 3.94\ fMF + 3.07\ IC0 - 4.72\ SIC0 - 0.02\ MIC4 - 0.17\ VSA\_EState8 + 0.04\ Wpol$
2. $\delta_p = 2.26 - 2.55\ nN - 3.27\ nO + 0.003\ ATS8v$ - 2. $26\ ATSC0c + 2.84\ ATSC2c + 0.02\ ATSC3dv - 0.07\ ATSC6p - 1.71\ BalabanJ + 0.46\ Xpc\text{-}4dv - 1.25\ AXp\text{-}1dv + 0.18\ SdO - 2.21\ nHBDon + 5.19\ IC0 + 0.94\ IC1 - 0.17\ Kier1$- $0.04\ PEOE\_VSA6 + 0.14\ SlogP\_VSA2 - 0.13\ VSA\_EState2 + 0.27\ TopoPSA(NO) - 8.50\ GGI5 + 0.04\ TSRW10$
3. $\delta_h = 11.59 + 0.44\ nHetero - 1.79\ nO - 0.04\ AATS0v + 1.79\ AATS0p - 4.93\ ATSC0c + 0.04\ ATSC0dv - 0.01\ ATSC7dv - 0.29\ ATSC3se + 144.59\ AATSC0c - 1.19\ AATSC1dv - 1.10\ AATSC0i - 2.37\ AXp\text{-}1dv - 0.79\ NsssCH + 2.07\ SsssCH + 0.48\ SsOH + 1.15\ nHBAcc + 0.89\ nHBDon + 1.74\ IC0 - 1.28\ CIC1 + 0.12\ EState\_VSA2 + 0.04\ EState\_VSA7 - 6.03\ GGI5$

Within the dispersive HSP component (δ_d), there are significant positive contributions of molecular framework fMF and information content index IC0, and negative contribution of topological descriptor SIC0 (Eq. (5a)). The information content index (IC) and its derivatives (order 0-2) are equal to the average information content ($\bar{IC}$) multiplied by the total number of atoms. The $\bar{IC}$ is based on Shannon's information theory, which provides information related to the effect of molecular symmetry within δ_d forces. For hydrogen-bond δ_h modeling, there is also a significant influence of information content indices (IC0, IC1). The molecular attributes that describe the average atomic charge distribution on the molecular topology (ATSC0c) showed a negative contribution to the polar HSP forces (δ_p). Within the δ_h HSP component, the GGI5 descriptor showed a negative contribution. The GGI5 descriptor is a topological charge index of order 5, proposed by Galvez et al. [70,71], and represents the total amount of charge transfer in the molecule. It suggests that the net charge transfer between five atoms, among others, mostly affects hydrogen bonding. ATSC0c is an autocorrelation descriptor calculated by taking the sum of the formal charges of atoms and is found to have a positive effect on δ_h forces. A higher ATSC0c value shows the influence of electronegative groups and hence, the charges imparted by these groups are favorable for hydrogen bonding.

So far, a detailed molecular description of Hansen solubility concerning individual HSP values (δ_d, δ_p, δ_h) has not been done. In the literature, there is an application of σ-moments and COSMO-RS descriptors for the prediction of Hansen solubility [72]. In ref. [34] Sanchez-Lengeling et al. select the influence of the shape and size of molecules, electrostatic forces, molecular structure, and their σ-profile. In our study, the molecular basis of individual Hansen solubility was described through MLR and XGBoost modeling.Regarding the existing ML studies, our manuscript, thanks to obtained MLR coefficients, investigates the fundamental concept of HSP and describes it in more detail.

3.3. Practical application

The most commonly evaluated phenomena in the pharmaceutical industry lie within the solubility interactions based on the 3D Hansen solubility concept. By understanding the fundamental characteristics of the HSP sphere, the GNN-predicted HSPs can be successfully used to solve problems in standard analytical and pharmaceutical analysis. In Hansen space, R_a is the Euclidean distance, and R_0 is the experimentally measured interaction radius. To visualize the Hansen solubility spheres for the investigated compounds, predicted from the developed models, Python's matplotlib library was used. According to Hansen space (Fig. 10 (a)), all compounds with a high degree of human skin permeation rate are placed at the center of the sphere with coordinates $\delta_h = 11$, $\delta_p = 12.5$, $\delta_d = 17.6$, and interaction radius $R_0 = 5$ [5]. The



(a) (b)

Fig 10. Applicability of GNN-predicted Hansen solubility based on HSP sphere method (a) permeation rate of selected solvents (b) lignin solubility.

GNN-predicted HSP values show that dimethyl sulfoxide is expected to have a higher degree of human skin permeation rate compared to n-butanol and cyclohexane.

Selecting the best solvent for a particular analytical purpose is also one of the most imperative concepts among the chemical analysis. By fixing the R_0 value and applying the RED formula to the predicted HSP values, it is possible to accurately report a chemical pair as soluble or insoluble. The interaction sphere of lignin [5] example considering the GNNs-predicted HSP values of the model solvents is given in Fig. 10 (b). All good and 'limiting' solvents are contained inside the sphere ($R_0 > R_a$) or at least on its surface ($R_0 = R_a$), closer to the geometrical center of the sphere. According to the results obtained in Fig. 10 (b), dimethyl sulfoxide can be classified as 'good'($R_0 > R_a$), n-butanol as a solvent of moderate type ($R_0 = R_a$), while the cyclohexane has 'poor' characteristics for dissolving lignin ($R_0 < R_a$).

From the obtained results (Table 2), it can be seen that there is a weaker agreement between the experimental data from the literature and the predicted values of the dispersive HSP component (R^2=0.66). It is known that dispersive forces can strongly depend on the shape and position of the molecule in space. Therefore, we might expect better results using 3D molecular interpretation compared to 2D and graph-based molecular interpretation in δ_d modeling.

4. Conclusions

Our results show that the three-dimensional Hansen solubility concept can be successfully predicted using Mordred descriptors (MLR, XGBoost) and graph-based molecular interpretation (GNN). The created computational models meet the criteria of internal and external validation metrics and can be used in the reliable HSP predictions. The fundamental basis of individual HSP forces has been successfully explained for MLR, and XGBoost10 modeling with the features associated with van der Waals surface area, 2D-matrix, and polarity. In addition, the coefficients of each parameter in the MLR equation revealed the specific influence of molecular interactions dominating Hansen solubility.

Our study also confirms the promising predictive capabilities of GNNs and their improved ability to generalize to unseen data. The best agreement between experimental HSPs from the literature and predicted data was found for GNN modeling including $R^2 > 0.65$ for dispersive and $R^2 > 0.75$ for polar and hydrogen bonding HSP forces. As such, the created models could enable solvent screening for chemical and pharmaceutical analysis.

Moving forward, future work would benefit from larger and more comprehensive datasets since both the quantity and quality of the data are of pivotal importance for machine learning in any field. Additionally, exploration into alternative models, varying combinations of GNN layers, or hybrid modeling could yield valuable insights. Beyond investigating alternative model approaches, incorporating additional features - particularly global molecular features, into GNN models could be a promising direction for future research. Furthermore, enhancing the interpretability of GNN models is of significant value. This can be achieved through existing methods for subgraph importance or through explainability techniques for the solute-solvent approach for HSP prediction. Improving interpretability not only increases trust in the models but could also provide additional physico-chemical insights into each Hansen solubility parameter. Nonetheless, our current research has introduced and proposed promising novel methods to predict and understand the concept of Hansen solubility.

Marija Mitrovic Dankulov: Writing – review & editing, Supervision, Investigation, Conceptualization. Aleksandar Bogojevic: Writing – review & editing, Supervision, Investigation, Conceptualization. Sasa Lazovic: Writing – review & editing, Supervision, Investigation, Conceptualization. Darija Obradovic: Writing – review & editing, Writing – original draft, Supervision, Investigation, Formal analysis, Data curation. Darja Cvetkovic: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Data curation

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared my data and code via the GitHub link in the paper, and in the Attach File step as Research Data.

Code and data for XGBoost and GNN prediction of Hansen Solubility Parameters (Original data) (GitHub)

Acknowledgements

The authors acknowledge funding provided by the Institute of Physics Belgrade,

National Institute of the Republic of Serbia through the grant from the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia. The authors would like to thank prof. Paola Gramatica on the free license of the software QSARINS.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105168.

References

[1] M. Belmares, M. Blanco, W.A. Goddard, R.B. Ross, G. Caldwell, S.-H. Chou, J. Pham, P.M. Olofson, C. Thomas, Hildebrand and Hansen solubility parameters from Molecular Dynamics with applications to electronic nose polymer sensors, J Comput Chem 25 (2004) 1814–1826, https://doi.org/10.1002/jcc.20098.

[2] P. Weerachanchai, Z. Chen, S.S.J. Leong, M.W. Chang, J.-M. Lee, Hildebrand solubility parameters of ionic liquids: Effects of ionic liquid type, temperature and DMA fraction in ionic liquid, Chemical Engineering Journal 213 (2012) 356–362, https://doi.org/10.1016/j.cej.2012.10.012.

[3] J.S. Delaney, ESOL: Estimating Aqueous Solubility Directly from Molecular Structure, J. Chem. Inf. Comput. Sci. 44 (2004) 1000–1005, https://doi.org/10.1021/ci034243x.

[4] A. Shayanfar, M.A.A. Fakhree, A. Jouyban, A simple QSPR model to predict aqueous solubility of drugs, Journal of Drug Delivery Science and Technology 20 (2010) 467–476, https://doi.org/10.1016/S1773-2247(10)50080-7.

[5] C.M. Hansen, Hansen Solubility Parameters: A User's Handbook, Second Edition, 0 ed., CRC Press, 2007. https://doi.org/10.1201/9781420006834.

[6] J. Henry Hildebrand, R. Lane Scott, The Solubility of Nonelectrolytes, Dover Publications, New York, 1964.

[7] J. Henry Hildebrand, Solubility of non-electrolytes. Von Prof. Joel H. Hildebrand. 203 Seiten. Reinhold Publishing Corporation, New York 1936, Seiten. Reinhold Publishing Corporation (1936). https://doi.org/10.1002/ange.19360493815.

[8] J.C. Zuaznabar-Gardona, A. Fragoso, Determination of the Hansen solubility parameters of carbon nano-onions and prediction of their dispersibility in organic solvents, Journal of Molecular Liquids 294 (2019) 111646, https://doi.org/10.1016/j.molliq.2019.111646.

[9] M.A. Mohammad, A. Alhalaweh, S.P. Velaga, Hansen solubility parameter as a tool to predict cocrystal formation, International Journal of Pharmaceutics 407 (2011) 63–71, https://doi.org/10.1016/j.ijpharm.2011.01.030.

[10] D. Obradović, F. Andrić, M. Zlatović, D. Agbaba, Modeling of Hansen's solubility parameters of aripiprazole, ziprasidone, and their impurities: A nonparametric comparison of models for prediction of drug absorption sites, Journal of Chemometrics 32 (2018) e2996, https://doi.org/10.1002/cem.2996.

[11] J. Ouyang, L. Liu, L. Zhou, Z. Liu, Y. Li, C. Zhang, Solubility, dissolution thermodynamics, Hansen solubility parameter and molecular simulation of 4-chlorobenzophenone with different solvents, Journal of Molecular Liquids 360 (2022) 119438, https://doi.org/10.1016/j.molliq.2022.119438.

[12] F.-J. Navarro-Lupión, P. Bustamante, B. Escalera, Relationship between swelling of hydroxypropylmethylcellulose and the Hansen and Karger partial solubility parameters, Journal of Pharmaceutical Sciences 94 (2005) 1608–1616, https://doi.org/10.1002/jps.20370.

[13] T. Ban, C.-L. Li, Q. Wang, Determination of the solubility parameter of allyl imidazolium-based ionic liquid using inverse gas chromatography and Hansen solubility parameter in practice, Journal of Molecular Liquids 271 (2018) 265–273, https://doi.org/10.1016/j.molliq.2018.08.095.

[14] Q. Wang, Y. Chen, L. Deng, J. Tang, Z. Zhang, Determination of the solubility parameter of ionic liquid 1-allyl-3-methylimidazolium chloride by inverse gas chromatography, Journal of Molecular Liquids 180 (2013) 135–138, https://doi.org/10.1016/j.molliq.2013.01.012.

[15] L. Zhao, Q. Wang, K. Ma, Solubility Parameter of Ionic Liquids: A Comparative Study of Inverse Gas Chromatography and Hansen Solubility Sphere, ACS Sustainable Chem. Eng. 7 (2019) 10544–10551, https://doi.org/10.1021/acssuschemeng.9b01093.

[16] P. Choi, T.A. Kavassalis, A. Rudin, Estimation of the three-dimensional solubility parameters of alkyl phenol ethoxylates using molecular dynamics, Journal of Colloid and Interface Science 150 (1992) 386–393, https://doi.org/10.1016/0021-9797(92)90208-4.

[17] T.A. Kavassalis, P. Choi, A. Rudin, The Calculation of 3D Solubility Parameters Using Molecular Models, Molecular Simulation 11 (1993) 229–241, https://doi.org/10.1080/08927029308022510.

[18] A.-G. Sicaire, M. Vian, F. Fine, F. Joffre, P. Carré, S. Tostain, F. Chemat, Alternative Bio-Based Solvents for Extraction of Fat and Oils: Solubility Prediction, Global Yield, Extraction Kinetics, Chemical Composition and Cost of Manufacturing, IJMS 16 (2015) 8430–8453, https://doi.org/10.3390/ijms16048430.

[19] N.R. Tummala, C. Sutton, S.G. Aziz, M.F. Toney, C. Risko, J.-L. Bredas, Effect of Solvent Additives on the Solution Aggregation of Phenyl-C $_{61}$ –Butyl Acid Methyl Ester (PCBM), Chem. Mater. 27 (2015) 8261–8272, https://doi.org/10.1021/acs.chemmater.5b03254.

[20] M. Williams, N.R. Tummala, S.G. Aziz, C. Risko, J.-L. Brédas, Influence of Molecular Shape on Solid-State Packing in Disordered PC $_{61}$ BM and PC $_{71}$ BM Fullerenes, J. Phys. Chem. Lett. 5 (2014) 3427–3433, https://doi.org/10.1021/jz501559q.

[21] F. Eckert, A. Klamt, Fast solvent screening via quantum chemistry: COSMO-RS approach, AIChE Journal 48 (2002) 369–385, https://doi.org/10.1002/aic.690480220.

[22] E. Stefanis, C. Panayiotou, Prediction of Hansen Solubility Parameters with a New Group-Contribution Method, Int J Thermophys 29 (2008) 568–585, https://doi.org/10.1007/s10765-008-0415-z.

[23] S. Abbott, C.M. Hansen, H. Yamamoto, Hansen Solubility Parameters in Practice – Complete with software, data, and examples, 5th ed., n.d. www.hansen-solubility.com.

[24] C. Panayiotou, Solubility parameter revisited: an equation-of-state approach for its estimation, Fluid Phase Equilibria 131 (1997) 21–35, https://doi.org/10.1016/S0378-3812(96)03221-9.

[25] E. Stefanis, C. Panayiotou, A new expanded solubility parameter approach, International Journal of Pharmaceutics 426 (2012) 29–43, https://doi.org/10.1016/j.ijpharm.2012.01.001.

[26] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, (2016). https://doi.org/10.48550/ARXIV.1603.02754.

[27] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81, https://doi.org/10.1016/j.aiopen.2021.01.001.

[28] R.P. Sheridan, W.M. Wang, A. Liaw, J. Ma, E.M. Gifford, Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships, J. Chem. Inf. Model. 56 (2016) 2353–2360, https://doi.org/10.1021/acs.jcim.6b00591.

[29] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. Van Hoesel, H. Schopmans, T. Sommer, P. Friederich, Graph neural networks for materials science and chemistry, Commun Mater 3 (2022) 93, https://doi.org/10.1038/s43246-022-00315-6.

[30] G. Panapitiya, M. Girard, A. Hollas, J. Sepulveda, V. Murugesan, W. Wang, E. Saldanha, Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction, ACS Omega 7 (2022) 15695–15710, https://doi.org/10.1021/acsomega.2c00642.

[31] Q. Yang, H. Ji, H. Lu, Z. Zhang, Prediction of Liquid Chromatographic Retention Time with Graph Neural Networks to Assist in Small Molecule Identification, Anal. Chem. 93 (2021) 2200–2206, https://doi.org/10.1021/acs.analchem.0c04071.

[32] D. Jiang, Z. Wu, C.-Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu, T. Hou, Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models, J Cheminform 13 (2021) 12, https://doi.org/10.1186/s13321-020-00479-8.

[33] J.D. Perea, S. Langner, M. Salvador, J. Kontos, G. Jarvas, F. Winkler, F. Machui, A. Görling, A. Dallos, T. Ameri, C.J. Brabec, Combined Computational Approach Based on Density Functional Theory and Artificial Neural Networks for Predicting The Solubility Parameters of Fullerenes, J. Phys. Chem. B 120 (2016) 4431–4438, https://doi.org/10.1021/acs.jpcb.6b00787.

[34] B. Sanchez-Lengeling, L.M. Roch, J.D. Perea, S. Langner, C.J. Brabec, A. Aspuru-Guzik, A Bayesian Approach to Predict Solubility Parameters, Advcd Theory and Sims 2 (2019) 1800069, https://doi.org/10.1002/adts.201800069.

[35] H. Feng, P. Zhang, W. Qin, W. Wang, H. Wang, Estimation of solubility of acid gases in ionic liquids using different machine learning methods, Journal of Molecular Liquids 349 (2022) 118413, https://doi.org/10.1016/j.molliq.2021.118413.

[36] M. Abdullah, K. Chellappan Lethesh, A.A.B. Baloch, M.O. Bamgbopa, Comparison of molecular and structural features towards prediction of ionic liquid ionic conductivity for electrochemical applications, Journal of Molecular Liquids 368 (2022) 120620, https://doi.org/10.1016/j.molliq.2022.120620.

[37] K. Low, M.L. Coote, E.I. Izgorodina, Explainable Solvation Free Energy Prediction Combining Graph Neural Networks with Chemical Intuition, J. Chem. Inf. Model. 62 (2022) 5457–5470, https://doi.org/10.1021/acs.jcim.2c01013.

[38] S. Lee, M. Lee, K.-W. Gyak, S.D. Kim, M.-J. Kim, K. Min, Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks, ACS Omega 7 (2022) 12268–12277, https://doi.org/10.1021/acsomega.2c00697.

[39] D. Wu, Z. Zhu, J. Zhang, H. Wen, S. Jin, W. Shen, An Interpretable Solute–Solvent Interactive Attention Module Intensified Graph-Learning Architecture toward Enhancing the Prediction Accuracy of an Infinite Dilution Activity Coefficient, Ind. Eng. Chem. Res. 63 (2024) 8741–8750, https://doi.org/10.1021/acs.iecr.4c00107.

[40] J. Pang, A.W.R. Pine, A. Sulemana, Using natural language processing (NLP)-inspired molecular embedding approach to predict Hansen solubility parameters, Digital Discovery 3 (2024) 145–154, https://doi.org/10.1039/D3DD00119A.

[41] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, H. Kabir, Comparative Studies on Some Metrics for External Validation of QSPR Models, J. Chem. Inf. Model. 52 (2012) 396–408, https://doi.org/10.1021/ci200520g.

[42] Y. Agata, H. Yamamoto, Determination of Hansen solubility parameters of ionic liquids using double-sphere type of Hansen solubility sphere method, Chemical Physics 513 (2018) 165–173, https://doi.org/10.1016/j.chemphys.2018.04.021.

[43] C.M. Hansen, A.L. Smith, Using Hansen solubility parameters to correlate solubility of C60 fullerene in organic solvents and in polymers, Carbon 42 (2004) 1591–1597, https://doi.org/10.1016/j.carbon.2004.02.011.

[44] H. Moriwaki, Y.-S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, J Cheminform 10 (2018) 4, https://doi.org/10.1186/s13321-018-0258-y.

[45] B. Ramsundar, P. Eastman, P. Walters, V. Pande, Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more, First edition, revision, second release, O'Reilly, Beijing Boston Farnham Sebastopol Tokyo, 2021.
[46] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, Molecular graph convolutions: moving beyond fingerprints, J Comput Aided Mol Des 30 (2016) 595–608, https://doi.org/10.1007/s10822-016-9938-8.
[47] S. Ulenberg, K. Ciura, P. Georgiev, M. Pastewska, G. Ślifirski, M. Król, F. Herold, T. Bączek, Use of biomimetic chromatography and in vitro assay to develop predictive GA-MLR model for use in drug-property prediction among anti-depressant drug candidates, Microchemical Journal 175 (2022) 107183, https://doi.org/10.1016/j.microc.2022.107183.
[48] P. Gramatica, Principles of QSAR Modeling: Comments and Suggestions From Personal Experience, International Journal of Quantitative Structure-Property Relationships 5 (2020) 61–97, https://doi.org/10.4018/IJQSPR.20200701.oa1.
[49] K. Roy, I. Mitra, On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design, CCHTS 14 (2011) 450–474, https://doi.org/10.2174/138620711795767893.
[50] R.L. Haupt, S.E. Haupt, Practical Genetic Algorithms, 1st ed., Wiley, 2003. https://doi.org/10.1002/0471671746.
[51] B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods, Chemometrics and Intelligent Laboratory Systems 64 (2002) 91–99, https://doi.org/10.1016/S0169-7439(02)00068-0.
[52] M. Salari, M.H. Dehghani, A. Azari, M.D. Motevalli, A. Shabanloo, I. Ali, High performance removal of phenol from aqueous solution by magnetic chitosan based on response surface methodology and genetic algorithm, Journal of Molecular Liquids 285 (2019) 146–157, https://doi.org/10.1016/j.molliq.2019.04.065.
[53] I. Mehraein, S. Riahi, The QSPR models to predict the solubility of CO 2 in ionic liquids based on least-squares support vector machines and genetic algorithm-multi linear regression, Journal of Molecular Liquids 225 (2017) 521–530, https://doi.org/10.1016/j.molliq.2016.10.133.
[54] Z.-Y. Yang, Z.-J. Yang, J. Dong, L.-L. Wang, L.-X. Zhang, J.-J. Ding, X.-Q. Ding, A.-P. Lu, T.-J. Hou, D.-S. Cao, Structural Analysis and Identification of Colloidal Aggregators in Drug Discovery, J. Chem. Inf. Model. 59 (2019) 3714–3726, https://doi.org/10.1021/acs.jcim.9b00541.
[55] T. Lei, H. Sun, Y. Kang, F. Zhu, H. Liu, W. Zhou, Z. Wang, D. Li, Y. Li, T. Hou, ADMET Evaluation in Drug Discovery. 18. Reliable Prediction of Chemical-Induced Urinary Tract Toxicity by Boosting Machine Learning Approaches, Mol. Pharmaceutics 14 (2017) 3935–3953, https://doi.org/10.1021/acs.molpharmaceut.7b00631.
[56] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, D.D. Cox, Hyperopt: a Python library for model selection and hyperparameter optimization, Comput. Sci. Disc. 8 (2015) 014008, https://doi.org/10.1088/1749-4699/8/1/014008.
[57] T.N. Kipf, M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, (2016). https://doi.org/10.48550/ARXIV.1609.02907.
[58] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph Attention Networks, (2017). https://doi.org/10.48550/ARXIV.1710.10903.
[59] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, M. Zheng, Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism, J. Med. Chem. 63 (2020) 8749–8760, https://doi.org/10.1021/acs.jmedchem.9b00959.
[60] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural Message Passing for Quantum Chemistry, (2017). https://doi.org/10.48550/ARXIV.1704.01212.
[61] S. Jaeger, S. Fulle, S. Turk, Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition, J. Chem. Inf. Model. 58 (2018) 27–35, https://doi.org/10.1021/acs.jcim.7b00616.
[62] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, (2020). https://doi.org/10.48550/ARXIV.2010.09885.
[63] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa-2: Towards Chemical Foundation Models, (2022). https://doi.org/10.48550/ARXIV.2209.01712.
[64] Notes on the N-Person Game — II: The Value of an N-Person Game, RAND Corporation, 1951. https://doi.org/10.7249/RM0670.
[65] B. Hollas, Autocorrelation Descriptor for Molecules, Journal of Mathematical Chemistry 33 (2003) 91–101, https://doi.org/10.1023/A:1023247831238.
[66] J.W. Godden, J. Bajorath, Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors, J. Chem. Inf. Comput. Sci. 41 (2001) 1060–1066, https://doi.org/10.1021/ci0102867.
[67] P. Nguyen, D. Loveland, J.T. Kim, P. Karande, A.M. Hiszpanski, T.Y.-J. Han, Predicting Energetics Materials' Crystalline Density from Chemical Structure by Machine Learning, J. Chem. Inf. Model. 61 (2021) 2147–2158, https://doi.org/10.1021/acs.jcim.0c01318.
[68] K. Roy, G. Ghosh, Exploring QSARs with Extended Topochemical Atom (ETA) Indices for Modeling Chemical and Drug Toxicity, CPD 16 (2010) 2625–2639, https://doi.org/10.2174/138161210792389270.
[69] M. Oja, S. Sild, U. Maran, Logistic Classification Models for pH–Permeability Profile: Predicting Permeability Classes for the Biopharmaceutical Classification System, J. Chem. Inf. Model. 59 (2019) 2442–2455, https://doi.org/10.1021/acs.jcim.8b00833.
[70] L. Bertato, N. Chirico, E. Papa, QSAR Models for the Prediction of Dietary Biomagnification Factor in Fish, Toxics 11 (2023) 209, https://doi.org/10.3390/toxics11030209.
[71] J. Galvez, R. Garcia, M.T. Salabert, R. Soler, Charge Indexes. New Topological Descriptors, J. Chem. Inf. Comput. Sci. 34 (1994) 520–525, https://doi.org/10.1021/ci00019a008.
[72] J.P. Wojeicchowski, A.M. Ferreira, T. Okura, M. Pinheiro Rolemberg, M.R. Mafra, J.A.P. Coutinho, Using COSMO-RS to Predict Hansen Solubility Parameters, Ind. Eng. Chem. Res. 61 (2022) 15631–15638, https://doi.org/10.1021/acs.iecr.2c01592.
[73] J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence, Nat Mach Intell 2 (2020) 573–584, https://doi.org/10.1038/s42256-020-00236-4.
[74] Z. Wu, J. Wang, H. Du, D. Jiang, Y. Kang, D. Li, P. Pan, Y. Deng, D. Cao, C.-Y. Hsieh, T. Hou, Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking, Nat Commun 14 (2023) 2585, https://doi.org/10.1038/s41467-023-38192-3.
[75] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, T. Unterthiner, Interpretable Deep Learning in Drug Discovery, (2019). http://arxiv.org/abs/1903.02788.
[76] Q. Yang, H. Ji, X. Fan, Z. Zhang, H. Lu, Retention time prediction in hydrophilic interaction liquid chromatography with graph neural network and transfer learning, Journal of Chromatography A 1656 (2021) 462536, https://doi.org/10.1016/j.chroma.2021.462536.
[77] J.K. Weber, J.A. Morrone, S. Bagchi, J.D.E. Pabon, S. Kang, L. Zhang, W.D. Cornell, Simplified, interpretable graph convolutional neural networks for small molecule activity prediction, J Comput Aided Mol Des 36 (2022) 391–404.
[78] J. Cremer, L. Medrano Sandonas, A. Tkatchenko, D.-A. Clevert, G. De Fabritiis, Equivariant Graph Neural Networks for Toxicity Prediction, Chem. Res. Toxicol. (2023), https://doi.org/10.1021/acs.chemrestox.3c00032.