Neural Mulliken Analysis: Molecular Graphs from Density Matrices for QSPR on Raw Quantum-Chemical Data

Oleg I. Gromov
Email: aalchm@gmail.com ORCID: 0000-0002-4119-8602

December 2, 2024

Abstract

Since the introduction of Graph Neural Networks (GNNs), molecular graphs have become useful tools in chemical informatics. However, in property prediction tasks, graph embeddings often still resemble traditional fingerprints. Here, we propose a straightforward approach to provide modern GNNs with raw quantum-chemical data, enabling efficient solutions to a range of chemical machine-learning problems.

The central role is played by the 1-electron density matrix derived from quantum chemical calculations (e.g. Hartree-Fock, DFT). The diagonal blocks of the density matrix are used as embeddings for the atomic nodes (*"atoms"*) in the molecular graph. Unlike conventional molecular graph representations, the chemical bond concept is not used. Instead, an additional set of nodes (*"links"*) between pairs of atoms is introduced. Their embeddings are the off-diagonal blocks of the density matrix, related to particular atom pairs. Directed graph edges connect either *"atoms"* with *"links"* or *vice versa*. The embeddings of the edges are derived from the basis set overlap matrix. The overlaps serve two purposes: first, they encode structural information such as distances and angles. Second, they act as weights in pooling operations. The use of element-wise multiplication of densities and overlaps is inspired by the Mulliken population analysis scheme.

The proposed concept was further tested using the Solubility Challenge (2008) by Llinàs *et al.* (DOI: 10.1021/ci800058v). A GNN was trained on a small dataset comprising 94 aqueous solubilities of drug-like molecules and subsequently used to predict the aqueous solubilities of 28 test molecules. The model achieved an $RMSE$ of 0.68 and an R^2 of 0.76, outperforming all methods proposed at that time. In our view, this represents a promising approach, particularly considering that even in a preliminary test the proposed architecture seems to be able to achieve state-of-the-art accuracy.

1 Introduction

Over the past decades, advancements in machine and deep learning (ML&DL) fields, along with improvements in hardware capabilities, have been rapidly adopted across various areas of science and technology. These developments have given rise to entirely new domains and significantly contributed to existing fields, such as cheminformatics[1, 2]. In chemistry, machine learning methods are now widely applied in diverse areas, ranging from drug design[3] to materials science[4] and from molecular simulations[5, 6] to self-driving laboratories[7].

A foundational problem in cheminformatics and related sub-fields is how to represent a molecule or molecular-level structure (e.g. polymers, crystals) in a machine-readable format suitable for applications such as machine learning (ML). Various options are available, including representing a molecular system as a feature vector (e.g. machine-learned fingerprints), a string (e.g. in the SMILES format), or a graph. In the ML context, the choice of molecular representation largely determines the applicable ML methods (e.g. due to data format constraints), the application domains (such as predictive or generative tasks), and the potential performance of the resulting ML models.

In recent years, the volume of research on molecular representations and their applications has grown so substantially that even a brief overview is both impractical and unnecessary here. Comprehensive reviews on molecular representation approaches are available elsewhere[8–10]. To show the place of the current work within the broader context and compare it to the state of the art, we have chosen a specific yet representative task from the QSPR domain: predicting the aqueous solubility of small molecules. This focused approach facilitates tracking recent advancements and conducting an indicative performance evaluation.

Predicting aqueous solubility is challenging[11] but critically important for applications such as drug design[12]. Over the past two decades, the accuracy of aqueous solubility predictions has seemingly improved gradually. As dataset sizes have grown from hundreds of molecules[13] to tens of thousands[14], new molecular representations and models, including molecular graphs and graph neural networks (GNNs), have been introduced. The best reported $RMSE$ (root-mean-square error) for predicting the $\log_{10} S$ (aqueous solubility decimal logarithm) has decreased from the 0.6–0.8 range in 2000–2001[15–17] to approximately 0.38, achieved by consensus models[18] in the EUOS/SLAS challenge[14]. A comprehensive review by Marcou *et al.*[19] provides an overview of these developments.

However, examining reported performance metrics, such as those compiled by Marcou *et al.*[19], does not clarify the current state of the field - in fact, it may obscure it. As noted by Marcou *et al.*, most studies report metrics derived from cross-validation, and the lack of standard external validation makes it difficult to assess real progress. This probably motivated Llinàs *et al.*[20]

to launch the first Solubility Challenge (SC-1) in 2008, which used predefined training and test sets of drug-like molecules with consistent solubility measurements. SC-1 created a unique opportunity to evaluate the state of the art under uniform conditions, with the best $RMSE$ reaching approximately 0.8[21]. It also sparked discussions about the factors limiting solubility prediction accuracy.

A decade later, as the lack of a widely accepted benchmark or consensus on data and prediction methods persisted, Llinàs and Avdeef[22] organized the second Solubility Challenge (SC-2) in 2019. This challenge focused heavily on ensuring the accuracy of experimental test data, resulting in *"tight"* and *"loose"* test datasets characterized by high (SD $\sim$ 0.17 log) and lower (SD $\sim$ 0.62 log) inter-laboratory reproducibility. Despite a bias toward drug-like molecules, these datasets remain among the most reliable external benchmarks for data accuracy and consistency. Surprisingly, SC-2 revealed little or no improvement in prediction accuracy compared to SC-1[11], despite significant advancements in molecular descriptors, the adoption of powerful techniques like boosting, and the rise of GNNs[23, 24], which use molecular graphs. In contrast, SC-1 predominantly employed methods like Random Forest Regression (RFR) and Multiple Linear Regression (MLR). Subsequent systematic comparisons of modern ML techniques, including various GNN architectures, generally confirm these findings. For instance, the best $RMSE$ achieved on the *"tight"* (easier) SC-2 test set was 0.7 (R^2 0.59), attained by a Graph Convolutional Neural Network (GCN)[25]. For the *"loose"* (harder) SC-2 test set the same model achieved an $RMSE$ of 1.62 (R^2 0.35).

A potential approach to improving the predictive capabilities of ML models is to provide them with more fundamental and physically grounded molecular representations. In our previous study[26], the direct use of raw electron density as input to a neural network for predicting the solubility of gases in polymers demonstrated promising results, achieving near-experimental accuracy with a dataset of only $\sim$200 chemical compounds. In that work, the chemical entities were represented by their real-space 3D electron density distributions, which were processed by a convolutional neural network (CNN). The relatively strong performance of the model was likely due to electron density providing comprehensive information about molecular properties, including spatial arrangement, charge distributions, and polarizability. However, this approach is not scalable and the computational workflow is resource-intensive. Consequently, only a few studies have exploited it to date[26–30].

Another conventional method for representing electron density and quantum systems involves the density matrix formalism, which is central to quantum mechanics. Analysis-wise, basic schemes such as Mulliken[31] and Löwdin[32] population analyses are commonly used to extract atomic charges from the density matrix. These charges often form the basis for force field parameterizations and molecular graph embeddings. In this paper, we propose a novel approach: directly converting the 1-electron density matrix of a molecule into a graph representation, thereby avoiding ambiguously defined entities such as atomic charges. Interestingly, a graph neural network (GNN) built upon this physically rigorous representation (limited only by the "exactness" of the underlying quantum chemical method) still retains some conceptual similarities to Mulliken analysis.

2 Definitions and Notations

To avoid overloading the text, we provide only the essential details necessary to understand the described approach, omitting formal definitions and derivations, which can be found elsewhere[33]. In this paper, we focus exclusively on the first-order, or equivalently, 1-electron density matrix, referred to hereafter as the density matrix for short.

Suppose the electronic wave function is expanded in a finite set of 1-electron basis functions $\{\phi_\mu(r)\}$, such as the widely used atomic orbital (AO) basis sets. The total electron charge density $\rho(r)$ at any point r can then be expressed as:

$$\rho(r) = \sum_{\mu\nu} P_{\mu\nu}\phi_\mu(r)\phi_\nu(r)$$

where $P_{\mu\nu}$ are the elements of the density matrix $\mathbf{P}$, which fully defines the charge density in a given basis set.

The total number of electrons, N, can be obtained by integrating $\rho(r)$:

$$N = \int \rho(r)dr = \sum_{\mu\nu} P_{\mu\nu} \int \phi_\mu(r)\phi_\nu(r)dr = \sum_{\mu\nu} P_{\mu\nu}S_{\mu\nu}$$

where $S_{\mu\nu}$ are the elements of the overlap matrix $\mathbf{S}$.

In this paper, we use AO-type basis sets, with the entire basis set $\{\phi_\mu(r)\}$ being composed of distinct subsets, each associated with a particular atom. Consequently, the matrices $\mathbf{P}$ and $\mathbf{S}$ can be partitioned into sets of blocks $\{\mathbf{P}_{\mu\nu}\}$ and $\{\mathbf{S}_{\mu\nu}\}$, where the u-th rowgroup and v-th colgroup correspond to the basis functions associated with the u-th and v-th atoms, respectively. It is important to note that commonly used basis sets are generally not orthogonal, and hence, the overlap matrix $\mathbf{S}$ is not an identity matrix. The flattened diagonal blocks $\{\mathbf{P}_{vv}\}$ will be denoted as $\rho_{A,v}$. The flattened off-diagonal blocks $\mathbf{P}_{uv}$ and $\mathbf{S}_{uv}$ will be denoted as $\rho_{L,(u,v)}$ and $s_{(u,v)}$, respectively.

3 Molecular Graph from Density Matrix

3.1 Graph

First, we define a node set *"atom"* ($\mathcal{A}$), corresponding to the atoms in a molecule. A diagonal block $\mathbf{P}_{vv}$ of the density matrix $\mathbf{P}$ expanded exclusively in a subset of basis functions associated with a v-th atom, is flattened and used as an embedding $\rho_{A,v}$ for the corresponding node (Fig. 1, top left). Since a $\mathbf{P}_{vv}$ block is symmetric, duplicate values can be omitted. The order of flattening

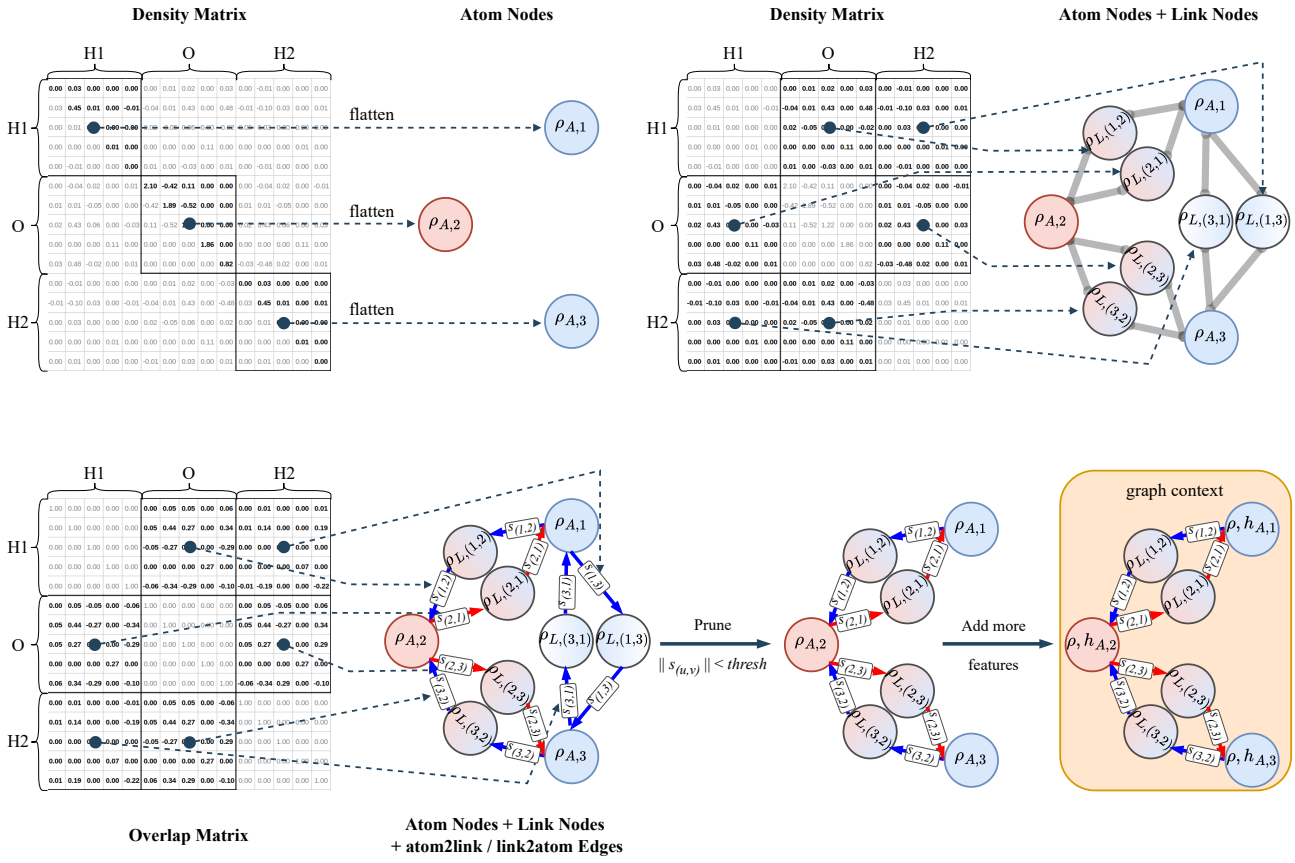

Figure 1: Construction of a molecular graph from a density matrix. H_2O molecule as an example. See Section 3.1 for a step-by-step explanation.

operation is arbitrary, provided it is consistent across the whole graph.

Next, we define a node set *"link"* ($\mathcal{L}$), which represents the electron density corresponding to pairs of atoms. Note that this node set is not necessarily equivalent to, for example, the set of chemical bonds in a molecule, although it does include them. An off-diagonal block $\mathbf{P}_{uv}$ of the density matrix $\mathbf{P}$, associated with the u-th and v-th atoms, is flattened and used as the embedding $\rho_{L,(u,v)}$ (Fig. 1, top right) for the corresponding node. A technical issue with this definition is that, unlike the diagonal blocks of the $\mathbf{P}$ matrix, the off-diagonal blocks are not symmetrical. The row and column indexes run over different subsets of basis functions, meaning that the order in which the matrix is flattened affects the result. To address this, we introduce two *"link"* nodes for every pair of atoms. Note that their embeddings, $\rho_{L,(u,v)}$ and $\rho_{L,(v,u)}$, derived from the off-diagonal blocks $\mathbf{P}_{uv}$ and $\mathbf{P}_{vu}$, respectively, are identical up to transposition. While this may seem redundant, it will be useful in the model we build on this graph.

Correspondingly, we define two sets of directed edges: *"atom2link"* ($\mathcal{A}2\mathcal{L}$) and *"link2atom"* ($\mathcal{L}2\mathcal{A}$). The direction of each edge encodes the order in which the off-diagonal blocks are flattened. At first glance, the introduction of $\mathcal{L}$, $\mathcal{A}2\mathcal{L}$, and $\mathcal{L}2\mathcal{A}$ instead of a single edge set may again seem redundant. However, this approach ensures that similar objects correspond to similar physical entities. Moreover, it will simplify the implementation of the model later on.

The graph defined in this way effectively encodes the electronic structure of a molecule and could potentially be used directly to train a graph neural network. However, two important aspects remain unaddressed. First, structural information, such as distances and angles, has not yet been incorporated. Second, the definition of the electron density distribution is incomplete without specifying the basis set.

An intuitive solution is to include additional embeddings derived from the basis set overlap matrix $\mathbf{S}$, using the same approach as with the $\mathbf{P}$ matrix (Fig. 1, bottom left). On the one hand, structural information is inherently encoded in the overlaps between basis functions. On the other hand, this ensures that all the data required to compute the actual electronic charges is included.

As a side note, it is perfectly possible to choose an orthonormal basis for a single atom. In such cases, the diagonal $\mathbf{S}_{vv}$ blocks are converted into identity matrices, leaving only the off-diagonal $\mathbf{S}_{uv}$ blocks to serve as edge embeddings ($s_{(u,v)}$).

Finally, some $\mathcal{L}$ nodes will naturally be removed because their corresponding edge embeddings are exactly zero due to zero basis overlap (a limitation of the precision in quantum chemical calculations). However, in general, the $\mathcal{L}$ set in the resulting (almost) fully connected molecular graph may be slightly excessive and could lead to unnecessary computations. Therefore, it is reasonable to remove links between very distant atoms.

In this paper, we prune the $\mathcal{L}$ set based on the

max norm of the corresponding overlap matrix blocks (Fig. 1, bottom right), which effectively serves as a simple distance-based filtering method. Further discussion can be found in Section 5.1.

The last technical detail to address is the selection of a suitable AO-basis set for expanding the **P** matrix. In conventional quantum chemistry, basis sets are typically tailored individually for each atom type or element. However, this approach is inconvenient in a machine learning context, as the elements of the **P** and **S** matrices become meaningless without explicit specification of the corresponding basis functions.

Moreover, it is preferable to assign an equal number of basis functions to every atom, ensuring that all node embeddings have equal dimensions without requiring additional padding. To achieve this, we propose assigning a unified, preferably orthonormal, AO-basis set to each atom (see Section 5.1).

3.2 Model

Given the molecular graph architecture described above, implementing a graph neural network is straightforward. We begin with the initialization of *"atom"* embeddings:

$$\rho^{(0)}_{A,v} = \rho_{A,v} || h_{A,v}$$

where $h_{A,v}$ represents additional features (e.g., nuclear masses, charges, etc.), and *"link"* embeddings:

$$\rho^{(0)}_{L,(u,v)} = \rho_{L,(u,v)}$$

Following the conventional approach, the node embeddings are updated iteratively:

$$\rho^{(k)}_{A,v}, \rho^{(k)}_{L,(u,v)} = GraphUpd(\rho^{(k-1)}_{A,v}, \rho^{(k-1)}_{L,(u,v)}, s_{(u,v)})$$

In our implementation, the message-passing layer first updates the *"link"* embeddings (Fig. 2):

$$\rho^{(k)}_{L,(u,v)} = MsgFn^{(k)}_{(A,L)}(\rho^{(k-1)}_{A,u} || s_{(u,v)} || \rho^{(k-1)}_{L,(u,v)}), \forall (u,v) \in G$$

where $MsgFn^{(k)}_{(A,L)}$ is a learnable function (e.g., a small neural network) shared across all nodes. Next, the *"atom"* embeddings are updated:

$$m^{(k)}_{(u,v)} = MsgFn^{(k)}_{(L,A)}(\rho^{(k)}_{L,(u,v)} || s_{(u,v)} || \rho^{(k-1)}_{A,v}), \forall (u,v) \in G$$
$$m^{(k)}_{v} = \sum_{u} m^{(k)}_{(u,v)} \odot s_{(u,v)}, \forall (u,v) \in G$$
$$\rho^{(k)}_{A,v} = MsgFn^{(k)}_{A}(m^{(k)}_{v} || \rho^{(k-1)}_{A,v})$$

Here, $MsgFn^{(k)}_{(L,A)}$ and $MsgFn^{(k)}_{A}$ are again learnable functions shared across all nodes. The messages $m^{(k)}_{v}$ received by the v-th *"atom"* node are computed as a weighted sum of the individual messages $m^{(k)}_{(u,v)}$ from the corresponding *"link"* nodes. The choice of whether to update the *"atom"* or *"link"* nodes first is arbitrary and can be determined through experimentation.

If the objective is to make graph-level predictions, the embeddings of *"atom"* nodes (and optionally, *"link"*

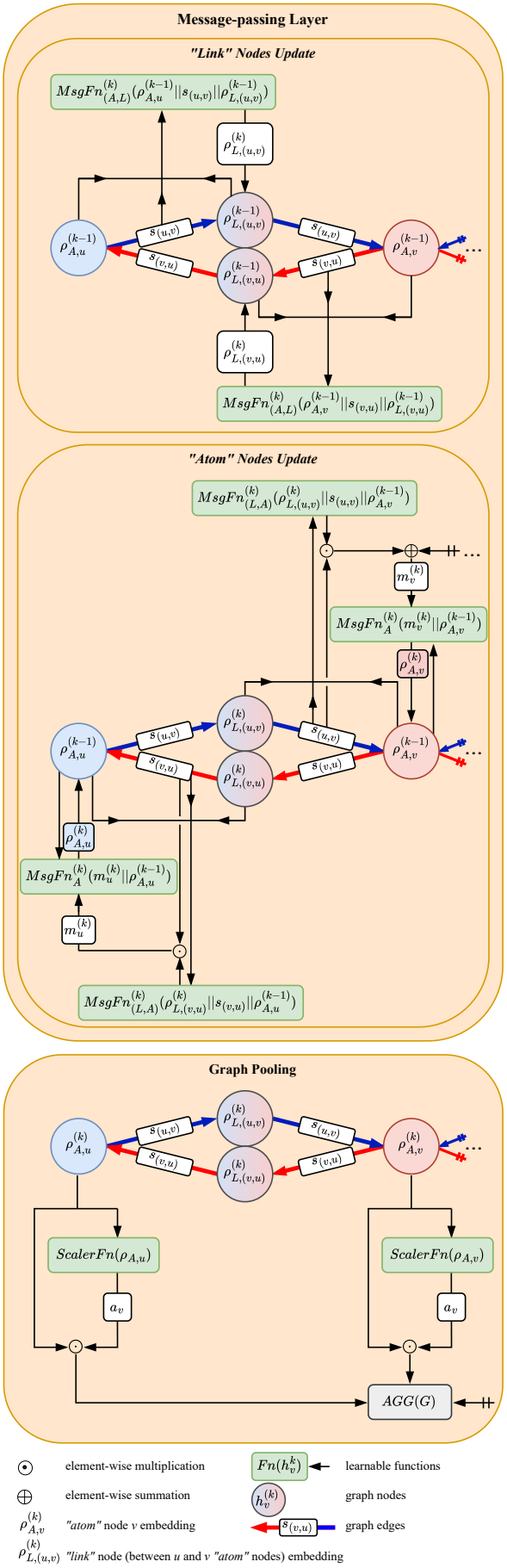

Figure 2: Message-passing and pooling layers.

nodes) are gathered and aggregated by a pooling layer to obtain a vector representation of the molecule:

$$\rho_G = PoolingLayer(\{\rho_{A,v}^{(k)}\}, \{\rho_{L,(u,v)}^{(k)}\})$$

In our implementation, node embeddings are first scaled according to their importance and then aggregated using element-wise summation, averaging, max, or a combination of these operations:

$$a_v = ScalerFn(\rho_{A,v}^{(k)}, \dots)$$
$$\rho_{A,v}^{(out)} = \rho_{A,v}^{(k)} \odot a_v$$
$$\rho_G = \overline{\rho_{A,v}^{(out)}} || \sum_G \rho_{A,v}^{(out)} || \dots$$

Here, *ScalerFn* is a learnable function shared across all nodes, which may optionally incorporate graph-level context. Essentially, this can be seen as a simplified form of attention. Finally, predictions are made:

$$\hat{y}_G = Predict(\rho_G)$$

where *Predict* is a learnable function, such as a linear layer or a multilayer perceptron.

Overall, aside from the message-passing mechanics, the model tested in this paper is a typical regression model: sequential message-passing steps are followed by graph pooling and a standard linear output layer. However, the described molecular graph architecture can also be effectively adapted for node-level tasks, which remains an interesting direction for further research.

4 Solubility Challenge

The Solubility Challenge (SC-1) by Llinàs et al. provides an excellent testbed for emerging QSPR techniques under challenging yet controlled conditions. The SC-1 dataset is toy-sized by modern standards, though this size may still reflect realistic scenarios in domains where data is both expensive and scarce. On the one hand, its compact size makes it convenient for quick performance evaluations, especially given its reliable and consistent experimental data and well-defined performance baseline for comparison. On the other hand, it may serve as a rigorous test of a model's ability to generalize from limited data, predict a complex property, and avoid overfitting.

The primary objective of this report was to demonstrate the potential of the proposed approach rather than to optimize the performance of a specific model for a specific task. Hence, we employed a small, heavily regularized 4-layer GNN (see Section 7.4 for details) with minimal hyperparameter tuning. In this configuration, an ensemble of 15 models achieved a test R^2 of 0.75 and $RMSE$ of 0.69, representing a clear improvement over the best R^2 of 0.65 and $RMSE$ of 0.80 reported in SC-1.

Another metric used in SC-1 is the percentage of *"correct"* predictions, defined as those with an absolute deviation from the measured value of less than 0.5 log units (denoted as $\pm 0.5\,\log$). In our case, the share of

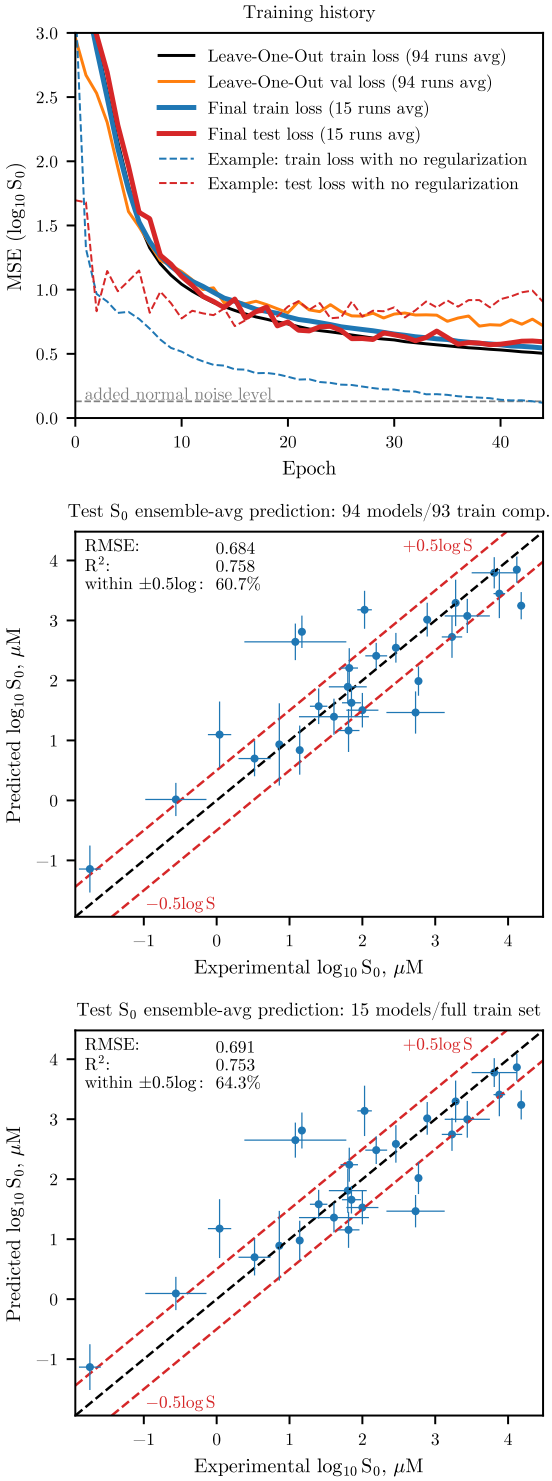

Figure 3: **Top**: Fitting loss. The *leave-one-out* (LOO) train (black) and val. (orange) loss - avgs. over 94 fits (93 train, 1 val. compound per fit). Final train (blue) and test (red) loss - avgs. over 15 fits (full train set). Test loss (28 test compounds) is for illustrative purposes. An example of a fit without regularization (train: blue dashed line, test: red dashed line) - a single run (full train and test sets). Grey dashed horizontal line - the level of noise added for regularization. **Middle** & **bottom**: Predicted *vs.* measured $\log_{10} S_0$ values of 28 test compounds. Middle: Averages of predictions of an ensemble of 94 models after LOO validation (trained on 93 compounds each). Bottom: Averages of predictions of an ensemble of 15 models (full training set). Error bars: horizontal - the standard deviations from the orig. challenge[21] or from inter-laboratory compilations[34, 35]; vertical - the standard deviations of predictions within ensemble.

"correct" predictions is 64.3%, compared to the best 60.7% reported in SC-1. However, we do not suppose this metric to be a reliable quantitative indicator due to the small size of the test set. For example, several borderline test compounds in our study exhibit errors fluctuating around 0.5 log units (Fig. 3), resulting in the share of *"correct"* predictions varying between 54% and 75% across individual models while the corresponding $RMSE$ are virtually identical.

SC-1 identified four particularly challenging compounds for solubility prediction: *indomethacin*, *terfenadine*, *diflunisal*, and *probenecid*. The originally reported solubility of *indomethacin* was later found to be erroneous[36]; using the updated value, our prediction deviates by an acceptable ~ 0.25 log units. For *terfenadine*, the average prediction error is also reasonable, at $\sim$0.5 log units. However, the solubilities of *diflunisal* and *probenecid* remain difficult to predict, with deviations of ~ 1.5 log units. This appears to stem from the limited flexibility of the applied models. Preliminary experiments with deeper and less regularized models occasionally reduced the prediction errors for *diflunisal* and *probenecid* to well below 1 log unit, but there were issues with instability and overfitting.

Overall, with additional data augmentation and thorough model and hyperparameter tuning, achieving close to 100% *"correct"* predictions appears feasible. However, the primary limitation to accuracy in such cases is the need to account for factors not solely determined by molecular structure. For instance, different polymorphs of the same compound can exhibit varying solubilities. Similarly, mixtures, such as those containing different enantiomers, may have solubilities significantly different from those of stereochemically pure compounds.

In the context of SC-1, these issues represent inherent limitations to the maximum achievable accuracy due to the small amount of data. However, in a broader context, with access to larger, more tailored datasets, these challenges seem to be more technical in nature (see Section 5.3 for further discussion). Regardless, further optimization of the challenge solution and understanding of accuracy limitations, while feasible, lies beyond the scope of this study.

5 Discussion

5.1 Graph and model details

The graph and model architectures presented in Section 3 offer significant flexibility in terms of the order of operations (e.g., update sequence), graph connectivity pruning methods, and context usage. An important note regarding our Solubility Challenge solution, described in Section 4, is that we consistently prioritized lower computational cost and ease of implementation. As such, it should be regarded as a quick illustrative example rather than a comprehensive case study. Consequently, the accuracy of our implementation is likely suboptimal.

In most cases, no systematic comparisons were performed when choosing between equivalent alternatives. For example, the decision to update *"link"* embeddings before *"atom"* embeddings, rather than the reverse order, was based on a few preliminary model fitting runs and was not revisited. The only exception was the selection of regularization parameters, which was determined using leave-one-out validation - still performed quite sketchy.

To filter out excessive *"link"* nodes, we applied a simple threshold of 0.035 for the maximum element of the overlap matrix, which roughly corresponds to a distance of $3\mathring{A}$. This ensures that every atom in a typical five- or six-membered ring is directly connected to all other atoms within the ring. No additional tests with alternative thresholds were performed. While simple, this strategy provides flexibility: it can yield either a more lightweight model that captures interactions with only the nearest neighbors or a more computationally intensive one that directly accumulates interactions within, for example, entire aromatic subsystems of a molecule in a single step rather than through a messaging chain. More sophisticated approaches could also be implemented, such as using a predefined set of conventional chemical bonds, introducing variable thresholds based on nuclear charge, or employing criteria based on natural bonding orbitals or van der Waals surface overlaps.

As mentioned in Section 3.1, the approach used in quantum chemistry to assign basis functions to atoms is not probably optimal for machine learning. In quantum chemistry, with a limited basis set size, different numbers of basis functions are typically assigned to different atom types to achieve a balanced description of molecular orbitals across the system. For instance, the popular def2-SVP basis set[37] assigns only five composite basis functions to hydrogen atoms (one electron) but 18 basis functions to chlorine atoms (17 electrons). Additionally, the basis sets for heavier atoms are not simply extensions of those for lighter atoms but are systematically tuned to achieve accuracy with a limited number of functions.

Directly using a unified basis set (as is optimal for ML applications) in quantum chemical calculations may result in an inaccurate description of the electronic structure. A possible workaround involves computing the electronic structure in a conventional basis set (e.g., def2-SVP, def2-TZVP) and projecting it onto a unified basis set, with identical subsets of basis functions assigned to all atoms. However, the resulting expansion will not be an *"optimal"* solution within the new basis set, and further convergence is avoided.

In this study, we used argon (Ar) basis functions from the minimal ANO-R0 basis set[38], where ANO stands for Atomic Natural Orbitals. These basis functions are constructed by fitting the natural orbitals of, for example, an Ar atom, forming an orthonormal basis. The minimal basis set was chosen to reduce embedding dimensions and computational effort. For greater accuracy, a larger basis set may be preferable.

Finally, certain model details were determined by the characteristics of the dataset and the property being predicted. Technically, the model consists of a GNN encoder and a $Predict()$ function. The encoder translates the input molecular graph into a latent space representation - a feature vector. The $Predict()$ function then

maps this feature vector to a property prediction.

Given the small dataset size, the prediction function might need sometimes to extrapolate on evaluation. Therefore, using a single linear layer at the output appears to be the most reliable option. On the other hand, as the number of chemical elements and atomic types in the dataset (and generally!) is relatively limited, the GNN encoder should benefit from a deeper architecture and non-linear message-passing functions. For the reported results, we used a 4-layer GNN, though we also experimented with deeper models (up to 16 layers). Occasionally, deeper models achieved even lower test $RMSE$ than those reported but exhibited increasing instability with depth. Further research on initialization strategies may be necessary, as switching to *LeCun* initialization somewhat improved model stability.

In this implementation, for all message functions ($MsgFn$) single-layer Kolmogorov-Arnold Networks (KAN)[39] are used. While KANs did not significantly improve accuracy, they provided faster convergence and enhanced stability to some extent, which makes them a convenient choice.

The choice of the node embedding aggregation method is tailored to the property being predicted. For instance, in solubility prediction, incorporating element-wise summation pooling is critical as it facilitates molecular volume prediction, a key factor correlated with solubility. In the present case, sum-pooling reduced the prediction error for *terfenadine* solubility (one of the toughest test compounds) by 2.5 log units compared to mean-pooling.

5.2 Data augmentation

The molecular graph representation presented here is not invariant, similar to the electron density images discussed in our previous paper[26]. Specifically, the embeddings of the molecular graph are expressed in a particular coordinate system; thus, when the coordinate system is rotated, the expansions of molecular orbitals and the density matrix change numerically. This issue, well-recognized in quantum chemistry, also affects population analysis based on the density matrix[40]. The underlying cause is that certain Gaussian basis functions, such as p_x, p_y, and p_z, lack spherical symmetry. To address this, the calculated data in this study has been extensively augmented with differently oriented copies of the same molecules. Nonetheless, developing an equivariant representation could be advantageous, although the method for achieving this is not immediately apparent.

Another type of invariance that is not accounted for is conformational invariance. Molecules can adopt multiple spatial arrangements of atoms, corresponding to different local minima on the potential energy surface. These configurations are known as conformers. The electronic structure of a molecule is inherently tied to a specific spatial arrangement of its atoms.

Given this, learning conformer-specific properties is relatively straightforward. However, if the goal is to predict a property related to the compound as a whole, such as solubility or reactivity, the choice of a specific molecular conformation for electronic structure calculations becomes challenging. Selecting a single conformer for each molecule, such as the lowest-energy conformer, may be both ambiguous and impractical. For example, when a process involves multiple physical phases (e.g., solid and liquid phases or different solvents), the lowest-energy conformer could vary between environments. Moreover, even with well-defined criteria for conformer selection, such choices may limit or complicate the applicability of the resulting model, particularly when data availability is restricted, such as for newly synthesized compounds.

The Solubility Challenge is an example of a task of predicting a property of a compound in general, which depends on its behaviour across several physical phases. To address the conformer selection issue, we adopt a strategy similar to that in our previous paper[26]. For each molecule in the dataset, we include multiple energetically accessible conformers, effectively expanding the dataset so that each original example gives rise to several new examples differing only in the molecular conformation used. This approach is intended to help the model learn general features while ignoring conformer-specific details.

It is worth noting that tautomeric isomerism presents an even more important aspect to account for[41]. As with conformational isomerism, this issue is handled by including multiple tautomers for a single chemical entity, as described in the Methods section.

5.3 Solubility Challenge

In the SC-1 context, avoiding overfitting is challenging due to the small training dataset and the high flexibility required to predict the target property. To address this, we employed several regularization techniques in addition to standard L2 regularization. These included the on-the-fly addition of normal noise to the training set target values and multitask learning. The multitask approach used additional properties (partially correlated with aqueous solubility) as prediction targets, making use of transfer learning to reduce overfitting and improve performance.

Notably, these additional properties were used as prediction targets rather than graph-level input features. This choice was made because some of these properties – such as molecular surface, volume, and dipole moment - are directly derived from quantum chemical calculations, making them redundant as graph-level features. Others, like the melting point (m.p.), indeed contain supplementary information that could potentially serve as useful features. However, in the SC-1 context, m.p. was treated as a target rather than a feature. This was because the dataset included only four compounds with multiple polymorphs, which is most likely insufficient to learn from. More broadly, whether to use such properties as features or targets depends on the model's objectives. For instance, polymorphs of novel, unsynthesized compounds are unlikely to be known, limiting their utility as features. Besides, with larger datasets, multitask learning as a regularization strategy may become unnecessary, but it can still be replaced by methods like conditional computation[42]/modular learning[43] to handle datasets with partially missing some of the target values.

As discussed in Section 4, mixtures (e.g., enantiomeric mixtures) present another factor limiting the highest possible accuracy, both in SC-1 and more generally. A potential solution is to employ a multi-tail model, where each component of a mixture is represented by a separate molecular graph input (with shared GNN-encoder). Enantiomeric mixtures would then become a special case of this broader approach.

A limitation of using SC-1 to evaluate performance is that direct comparisons are only possible with the original 2009 results. Modern research typically uses SC-2 test sets for external validation, if the external validation is used at all, of course. However, SC-1 and SC-2 share several characteristics: both challenges employed accurate solubility measurements of drug-like compounds using consistent methods, and the SC-1 test set partially overlaps both with the *"tight"* (easier) and *"loose"* (more challenging) test sets from SC-2. Given the similarities in reported results between SC-1 and SC-2, as well as the small size of the SC-1 training set, a model's performance in SC-1 conditions can be considered a reasonable lower bound for accuracy. Substantially better results can be expected with larger datasets and deeper architectures.

In the case of the presented architecture, it is evident that the model used for the SC-1 solution is overly regularized. Without regularization, the model is capable of minimizing training loss to the level of experimental noise. This suggests that the model would benefit significantly from a larger and more diverse dataset, potentially delivering even higher test accuracy.

6 Conclusions

In this paper, we introduced novel molecular graphs constructed from density matrices that enables QSPR techniques with raw quantum chemical data. These molecular graphs, while being compact and computationally efficient, provide access to accurate (within the limits of the underlying quantum chemical calculations) electronic structure of a molecule.

The proposed graph representation, coupled with the corresponding graph neural network, was evaluated under the challenging conditions of the Solubility Challenge (2008) and demonstrated highly competitive performance.

This approach is anticipated to be particularly well-suited for ADMET prediction tasks, especially in domains with limited data availability. Its ability to achieve enhanced accuracy and generalization on small training sets, albeit with slightly higher computational cost, makes it a promising option for such applications.

7 Methods

7.1 Calculation details

The calculation pipeline used in this research closely follows that employed in our previous paper[26] and uses the *qcdata_gen* routine from the related repository for conformer generation. For each molecule, up to 1000 conformers were generated using the ETKDGv2 method[44] and subsequently optimized using the MMFF94 force field[45] in RDKit[46]. Conformer geometries were then sequentially optimized using the GFN2-xTB[47] and B97-3c[48] methods in the ORCA[49] program package.

After each optimization step, high-energy conformers were filtered out using energy thresholds of 50 kJ/mol for MMFF94 and GFN2-xTB optimizations, and 10 kJ/mol for B97-3c optimization. The remaining conformers were then checked for duplicates using the *conformers* script[50] with a spatial proximity threshold $RMSD$ of $0.2\mathring{A}$. No special effort was made to ensure the completeness of the resulting conformer set or to capture specific conformers reported in the literature.

Electronic structures of the resulting conformers were initially calculated at the PBE/def2-SVP[37] level of theory. The resulting molecular orbitals were projected onto a unified basis set comprising identical subsets assigned uniformly to all atoms. Specifically, the minimal ANO-R0[38] basis set of the Ar atom was used. No further molecular orbitals optimization was performed. The resulting density matrices were used for constructing molecular graphs.

7.2 Molecular graph implementation details

The molecular graphs described in Section 3.1 were implemented using the TF-GNN library[51] within the TensorFlow2 framework[52]. The embeddings for *"atom"* nodes, *"link"* nodes, and edges ($\rho_{A,v}$, $\rho_{L,(u,v)}$, and $s_{(u,v)}$, respectively) are vectors of length 65, 81, and 81, as the electron density was represented in a basis set with 9 basis functions per atom. The embeddings of *"atom"* nodes were further augmented with the corresponding nuclear charges, which was found to slightly improve the model's performance.

7.3 Data update, data sources, and augmentation

For the architecture testing, the data from the original Solubility Challenge[20, 21] was used with slight modifications.

7.3.1 Compound selection

Certain compounds were excluded from the original training and test sets based on the following criteria:

- Compounds without measured intrinsic solubility were removed.

- Compounds containing fourth-row elements were replaced to maintain a small quantum chemical basis set. Specifically, *2-amino-5-bromobenzoic acid*, *amiodarone*, *5-bromo-2,4-dihydroxybenzoic acid*, *bromogramine*, and *4-iodophenol* were replaced with *aripiprazole*, *atovaquone*, *benzoic acid*, *deoxycholic acid*, and *tamoxifen*.

As a result, the revised training and test sets consisted of 94 and 28 unique compounds, respectively.

7.3.2 Solubility data update

The intrinsic solubility data was updated using inter-laboratory averages where available, based on data systematized by Avdeef[34, 35]:

- If multiple solubility measurements by the CheqSol[53] method were reported, the corresponding inter-lab average values were used[34].

- In cases where inter-lab reproducibility data for the CheqSol method were unavailable, but averages of values measured by other methods were provided[35], these averages replaced the original values.

- If no additional data was available, the original values were retained without further literature review.

This way we tried to improve the accuracy of the original dataset while maintaining consistency with its inherent bias, as the dataset initially relied exclusively on solubility measurements obtained using the CheqSol method.

7.3.3 Secondary targets

To enhance model performance, a set of additional targets partially correlated with solubility was included:

- **Melting temperature** and **octanol-water partition coefficient**: Experimental averages were obtained from CompTox[54]. If experimental values were unavailable, averages of calculated values were used without further literature search.

- **Molecular dipole moment magnitudes**: Values were taken directly from quantum chemical calculations (PBE/def2-SVP).

- **Molecular volumes** and **surfaces**: Calculations were performed using the GEPOL93 program[55] with van der Waals radii from Truhlar *et al.*[56]

No significant effort was made to improve the accuracy of the additional target values.

7.3.4 Target and input values scaling

The target values and quantum-chemical data were scaled as follows:

- **Intrinsic solubility** $S_0(\mu M)$ and the **octanol-water partition coefficient** K_{ow}: Transformed to decimal logarithms.

- **Melting temperature** (°C), **molecular volume** ($\mathring{A}^3$), and **surface area** ($\mathring{A}^2$): Scaled by a factor of 100.

- **Dipole moment magnitude** (D): Used without scaling.

- **Density** and **overlap matrix** values (a.u.): Used without scaling.

- **Nuclear charges** (a.u.): Scaled by a factor of 100.

7.3.5 Chemical structure selection

The compounds in the training set were also represented by several possible tautomeric isomers when (preferably experimental) case studies were available in the literature. Specifically, the following tautomeric forms were considered: *guanine* keto-enol tautomers[57], *azathioprine* tautomers[58, 59], *cyanoguanidine*[60, 61] and *cimetidines* imidazolyl[62] units tautomers, *tetracyclines* keto-enol tautomers[63], and *phenobarbital* keto-enol tautomers[64]. Tautomeric isomerism for amides was not considered automatically or indiscriminately[65]. Zwitterionic tautomers were accounted for in compounds such as *ciprofloxacin*, *norfloxacin*, *lomefloxacin*[66], *danofloxacin*[67], *tetracyclines*[68], and *piroxicam*[69]. Additionally, ring and chain tautomers were used for *warfarin*[70].

To ensure fair performance evaluation, no tautomeric isomers were used in the test set, as the intended application of machine learning methods often targets new compounds with no experimental data or prior tautomeric equilibrium studies.

For compounds with experimental measurements related to stereochemically pure forms, their mirror-symmetrical isomers were also included, given their identical properties. Augmenting the training set with tautomers and mirror-image structures resulted in 182 unique chemical entities.

Each isomeric structure was further represented by a set of conformers (see Section 5.2 for discussion). The number of possible conformers varied between chemical entities. In total, the training set included 1427 unique 3D structures, while the test set contained 328 unique 3D structures.

7.3.6 Rotational augmentation

To address the lack of rotational invariance in quantum chemical calculations, 50 randomly rotated electron density distributions were generated for each conformer from the training set. To check the models ability to generalize, 9 randomly rotated electron density distributions were used for each conformer from the test set.

7.3.7 Noisify data

On training, normal noise was added on-the-fly to the target values. To introduce physically meaningful variability in the quantum chemical data, approximately 10% of the training electron densities were calculated using the BLYP density functional instead of the default PBE. Additionally, around 10% of the training molecular graphs were pruned using a threshold of 0.19, corresponding to $\sim 2\mathring{A}$ (only bonded interactions retained).

7.4 Model and fitting details

The model consists of 4 message-passing layers, a pooling layer, and six single-layer linear output heads corresponding to each target (the primary intrinsic solubility target and five secondary targets). Within each message-passing layer, the node set embeddings $\rho_{A,v}$ and $\rho_{L,(u,v)}$ (densities) are updated as described in Section 3.2 (Fig. 2, top), with their dimensions (65 and 81, respectively) remaining constant. The edge sets' embeddings $s_{(u,v)}$ (overlaps) are static. All message-processing functions ($MsgFn()$) are represented by single-layer Kolmogorov-Arnold Networks (KANs)[39] with splines of order 2 and a grid size of 3. Tensorflow2 implementation of KANs by Zhou *et al.*[71] is used.

In the pooling layer (Section 3.2; Fig. 2, bottom), scaling coefficients for each *"atom"* node are computed using a single linear layer without activation ($ScalerFn()$).

Subsequently, standard sum-, max-, and mean-pooling operations are performed on the *"atom"* nodes. The resulting feature vectors are concatenated to produce the output of the GNN.

Models weights were initialized where applicable using the Normal LeCun[72] scheme as provided by the TensorFlow2 framework. The loss function was defined as the weighted sum of the mean squared errors (MSE) between the predicted and true target values. The target-specific weights were set as follows: 0.5 for $log_{10}S_0$, 1.5 for T_m, 0.5 for $log_{10}K_{ow}$, 5.0 for electric dipole moment, 1.5 for molecular volume, and 1.5 for molecular surface area. Optimization was performed using the AdamW optimizer[73] with a learning rate of 1×10^{-4}. L2 regularization was applied exclusively to the kernel, with a weight decay parameter λ of 0.8. Biases were not subject to regularization. To further mitigate overfitting, Gaussian noise was added to the target values on-the-fly during training. The standard deviations of the noise were as follows: 0.35 for $log_{10}S_0$, T_m, and $log_{10}K_{ow}$; and 0.1 for electric dipole moment, molecular volume, and molecular surface area. The minibatch size was fixed at 8 examples, and each epoch included 1000 steps. During each epoch, the model was fed an approximately equal number of examples corresponding to each compound. The specific electron densities (i.e., a compound's tautomer, conformer, and spatial orientation) were sampled randomly for each example.

The number of layers, regularization parameters, secondary target weights, and other hyperparameters were selected based on leave-one-out validation, as the dataset size was too small to allocate a dedicated validation subset. Predictions from a single model were averaged over its set of conformers and spatial orientations. Final predictions on the test set were obtained as consensus averages over predictions from 15 identically trained models.

Data availability statement

The code used in the present paper is available at github.com/Shorku/rhnet2

Corresponding author

Oleg I. Gromov; ORCID: 0000-0002-4119-8602; Email: aalchm@gmail.com

CRediT author statement

OIG: conceptualization, software, formal analysis, writing the original draft.

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 reported work.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] T. Engel. "Basic Overview of Chemoinformatics". In: *Journal of Chemical Information and Modeling* 46.6 (Nov. 2006), pp. 2267–2277. DOI: 10.1021/ci600234z.

[2] E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, and A. Tropsha. "QSAR without borders". In: *Chemical Society Reviews* 49.11 (2020), pp. 3525–3564. DOI: 10.1039/D0CS00098A.

[3] P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann, and G. Schneider. "Rethinking drug design in the artificial intelligence era". In: *Nature Reviews Drug Discovery* 19.5 (May 2020), pp. 353–364. DOI: 10.1038/s41573-019-0050-3.

[4] T. C. Le and D. A. Winkler. "Discovery and Optimization of Materials Using Evolutionary Approaches". In: *Chemical Reviews* 116.10 (May 2016), pp. 6107–6132. DOI: 10.1021/acs.chemrev.5b00691.

[5] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. "Machine learning of accurate energy-conserving molecular force fields". In: *Science Advances* 3.5 (May 2017). DOI: 10.1126/sciadv.1603015.

[6] P. B. Jørgensen and A. Bhowmik. "Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids". In: *npj Computational Materials* 8.1 (Aug. 2022), p. 183. DOI: 10.1038/s41524-022-00863-y. arXiv: 2112.00652.

[7] G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid, and A. Aspuru-Guzik. "Self-Driving Laboratories for Chemistry and Materials Science". In: *Chemical Reviews* 124.16 (Aug. 2024), pp. 9633–9732. DOI: 10.1021/acs.chemrev.4c00055.

[8] L. David, A. Thakkar, R. Mercado, and O. Engkvist. "Molecular representations in AI-driven drug discovery: a review and practical guide". In: *Journal of Cheminformatics* 12.1 (Dec. 2020), p. 56. DOI: 10.1186/s13321-020-00460-5.

[9] K. Atz, F. Grisoni, and G. Schneider. "Geometric deep learning on molecular representations". In: *Nature Machine Intelligence* 3.12 (Dec. 2021), pp. 1023–1032. DOI: 10.1038/s42256-021-00418-8.

[10] D. S. Wigh, J. M. Goodman, and A. A. Lapkin. "A review of molecular representation in the age of machine learning". In: *WIREs Computational Molecular Science* 12.5 (Sept. 2022). DOI: 10.1002/wcms.1603.

[11] A. Llinàs, I. Oprisiu, and A. Avdeef. "Findings of the Second Challenge to Predict Aqueous Solubility". In: *Journal of Chemical Information and Modeling* 60.10 (Oct. 2020), pp. 4791–4803. DOI: 10.1021/acs.jcim.0c00701.

[12] K. Balakin, N. Savchuk, and I. Tetko. "In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds: Trends, Problems and Solutions". In: *Current Medicinal Chemistry* 13.2 (Jan. 2006), pp. 223–241. DOI: 10.2174/092986706775197917.

[13] J. Huuskonen. "Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology". In: *Journal of Chemical Information and Computer Sciences* 40.3 (May 2000), pp. 773–777. DOI: 10.1021/ci9901338.

[14] A. Zaliani, J. Tang, J. Martin, R. Harmel, and W. Wang. *1st EUOS/SLAS joint challenge: compound solubility*. 2022. URL: https://kaggle.com/competit%0Aions/euos-slas.

[15] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, and A. E. P. Villa. "Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices". In: *Journal of Chemical Information and Computer Sciences* 41.6 (Nov. 2001), pp. 1488–1493. DOI: 10.1021/ci000392t.

[16] P. Bruneau. "Search for Predictive Generic Model of Aqueous Solubility Using Bayesian Neural Nets". In: *Journal of Chemical Information and Computer Sciences* 41.6 (Nov. 2001), pp. 1605–1616. DOI: 10.1021/ci010363y.

[17] R. Liu, H. Sun, and S.-S. So. "Development of Quantitative Structure-Property Relationship Models for Early ADME Evaluation in Drug Discovery. 2. Blood-Brain Barrier Penetration". In: *Journal of Chemical Information and Computer Sciences* 41.6 (Nov. 2001), pp. 1623–1632. DOI: 10.1021/ci010290i.

[18] A. Hunklinger, P. Hartog, M. Šícho, G. Godin, and I. V. Tetko. "The openOCHEM consensus model is the best-performing open-source predictive model in the First EUOS/SLAS joint compound solubility challenge". In: *SLAS Discovery* 29.2 (Mar. 2024), p. 100144. DOI: 10.1016/j.slasd.2024.01.005.

[19] P. Llompart, C. Minoletti, S. Baybekov, D. Horvath, G. Marcou, and A. Varnek. "Will we ever be able to accurately predict solubility?" In: *Scientific Data* 11.1 (Mar. 2024), p. 303. DOI: 10.1038/s41597-024-03105-6.

[20] A. Llinàs, R. C. Glen, and J. M. Goodman. "Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements?" In: *Journal of Chemical Information and Modeling* 48.7 (July 2008), pp. 1289–1303. DOI: 10.1021/ci800058v.

[21] A. J. Hopfinger, E. X. Esposito, A. Llinàs, R. C. Glen, and J. M. Goodman. "Findings of the Challenge To Predict Aqueous Solubility". In: *Journal of Chemical Information and Modeling* 49.1 (Jan. 2009), pp. 1–5. DOI: 10.1021/ci800436c.

[22] A. Llinàs and A. Avdeef. "Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ∼ 0.17 log) and Loose (SD ∼ 0.62 log) Test Sets". In: *Journal of Chemical Information and Modeling* 59.6 (June 2019), pp. 3036–3040. DOI: 10.1021/acs.jcim.9b00345.

[23] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[24] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. *Convolutional Networks on Graphs for Learning Molecular Fingerprints*. 2015. arXiv: 1509.09292 [cs.LG]. URL: https://arxiv.org/abs/1509.09292.

[25] T. Zheng, J. B. O. Mitchell, and S. Dobson. "Revisiting the Application of Machine Learning Approaches in Predicting Aqueous Solubility". In: *ACS Omega* 9.32 (Aug. 2024), pp. 35209–35222. DOI: 10.1021/acsomega.4c06163.

[26] O. I. Gromov. "Predicting the solubility of gases, vapors, and supercritical fluids in amorphous polymers from electron density using convolutional neural networks". In: *Polymer Chemistry* 15.13 (2024), pp. 1273–1296. DOI: 10.1039/D3PY01028G.

[27] S. Singh, G. Zeh, J. Freiherr, T. Bauer, I. Türkmen, and A. T. Grasskamp. "Classification of substances by health hazard using deep neural networks and molecular electron densities". In: *Journal of Cheminformatics* 16.1 (Apr. 2024), p. 45. DOI: 10.1186/s13321-024-00835-y.

[28] Y. Zhao, K. Yuan, Y. Liu, S.-Y. Louis, M. Hu, and J. Hu. "Predicting Elastic Properties of Materials from Electronic Charge Density Using 3D Deep Convolutional Neural Networks". In: *The Journal of Physical Chemistry C* 124.31 (Aug. 2020), pp. 17262–17273. DOI: 10.1021/acs.jpcc.0c02348.

[29] S. Kajita, N. Ohba, R. Jinnouchi, and R. Asahi. "A Universal 3D Voxel Descriptor for Solid-State Material Informatics with Deep Convolutional Neural Networks". In: *Scientific Reports* 7.1 (Dec. 2017), p. 16991. DOI: 10.1038/s41598-017-17299-w.

[30] A. D. Casey, S. F. Son, I. Bilionis, and B. C. Barnes. "Prediction of Energetic Material Properties from Electronic Structure Using 3D Convolutional Neural Networks". In: *Journal of Chemical Information and Modeling* 60.10 (Oct. 2020), pp. 4457–4473. DOI: 10.1021/acs.jcim.0c00259.

[31] R. S. Mulliken. "Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I". In: *The Journal of Chemical Physics* 23.10 (Oct. 1955), pp. 1833–1840. DOI: 10.1063/1.1740588.

[32] P.-O. Löwdin. "On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals". In: *The Journal of Chemical Physics* 18.3 (Mar. 1950), pp. 365–375. DOI: 10.1063/1.1747632.

[33] T. Helgaker, P. Jorgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. John Wiley & Sons, 2000, p. 938. ISBN: 978-0-471-96755-2.

[34] A. Avdeef. "Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods". In: *ADMET and DMPK* 7.3 (Aug. 2019), pp. 210–219. DOI: 10.5599/admet.698.

[35] A. Avdeef. "Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database". In: *ADMET and DMPK* 8.1 (Mar. 2020), pp. 29–77. DOI: 10.5599/admet.766.

[36] J. Comer, S. Judge, D. Matthews, L. Towers, B. Falcone, J. Goodman, and J. Dearden. "The intrinsic aqueous solubility of indomethacin". In: *ADMET & DMPK* 2.1 (Apr. 2014). DOI: 10.5599/admet.2.1.33.

[37] F. Weigend and R. Ahlrichs. "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy". In: *Physical Chemistry Chemical Physics* 7.18 (2005), p. 3297. DOI: 10.1039/b508541a.

[38] J. P. Zobel, P.-O. Widmark, and V. Veryazov. "The ANO-R Basis Set". In: *Journal of Chemical Theory and Computation* 16.1 (Jan. 2020), pp. 278–294. DOI: 10.1021/acs.jctc.9b00873.

[39] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark. *KAN: Kolmogorov-Arnold Networks*. 2024. arXiv: 2404.19756 [cs.LG]. URL: https://arxiv.org/abs/2404.19756.

[40] I. Mayer. "Löwdin population analysis is not rotationally invariant". In: *Chemical Physics Letters* 393.1-3 (July 2004), pp. 209–212. DOI: 10.1016/j.cplett.2004.06.031.

[41] Y. C. Martin. "Let's not forget tautomers". In: *Journal of Computer-Aided Molecular Design* 23.10 (Oct. 2009), p. 693. DOI: 10.1007/s10822-009-9303-2.

[42] E. Bengio, P.-L. Bacon, J. Pineau, and D. Precup. *Conditional Computation in Neural Networks for faster models*. 2016. arXiv: 1511.06297 [cs.LG]. URL: https://arxiv.org/abs/1511.06297.

[43] J. Pfeiffer, S. Ruder, I. Vulić, and E. M. Ponti. *Modular Deep Learning*. 2024. arXiv: 2302.11529 [cs.LG]. URL: https://arxiv.org/abs/2302.11529.

[44] S. Wang, J. Witek, G. A. Landrum, and S. Riniker. "Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences". In: *Journal of Chemical Information and Modeling* 60.4 (Apr. 2020), pp. 2044–2058. DOI: 10.1021/acs.jcim.0c00025.

[45] T. A. Halgren. "MMFF VI. MMFF94s option for energy minimization studies". In: *Journal of Computational Chemistry* 20.7 (May 1999), pp. 720–729. DOI: 10.1002/(SICI)1096-987X(199905)20:7<720::AID-JCC7>3.0.CO;2-X.

[46] G. Landrum. *RDKit: Open-source cheminformatics*. 2006. URL: http://rdkit.org/.

[47] C. Bannwarth, S. Ehlert, and S. Grimme. "GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions". In: *Journal of Chemical Theory and Computation* 15.3 (Mar. 2019), pp. 1652–1671. DOI: 10.1021/acs.jctc.8b01176.

[48] J. G. Brandenburg, C. Bannwarth, A. Hansen, and S. Grimme. "B97-3c: A revised low-cost variant of the B97-D density functional method". In: *The Journal of Chemical Physics* 148.6 (Feb. 2018). DOI: 10.1063/1.5012601.

[49] F. Neese, F. Wennmohs, U. Becker, and C. Riplinger. "The ORCA quantum chemistry program package". In: *The Journal of Chemical Physics* 152.22 (June 2020), p. 224108. DOI: 10.1063/5.0004608.

[50] A. Genaev. *conformers*. 2021. URL: http://limor1.nioch.nsc.ru/quant/program/conformers.

[51] O. Ferludin, A. Eigenwillig, M. Blais, D. Zelle, J. Pfeifer, A. Sanchez-Gonzalez, W. L. S. Li, S. Abu-El-Haija, P. Battaglia, N. Bulut, J. Halcrow, F. M. G. de Almeida, P. Gonnet, L. Jiang, P. Kothari, S. Lattanzi, A. Linhares, B. Mayer, V. Mirrokni, J. Palowitch, M. Paradkar, J. She, A. Tsitsulin, K. Villela, L. Wang, D. Wong, and B. Perozzi. *TF-GNN: Graph Neural Networks in TensorFlow*. 2023. arXiv: 2207.03522 [cs.LG]. URL: https://arxiv.org/abs/2207.03522.

[52] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. *TensorFlow: Large-Scale*

Machine Learning on Heterogeneous Distributed Systems. 2016. arXiv: 1603.04467 [cs.DC]. URL: https://arxiv.org/abs/1603.04467.

[53] M. Stuart and K. Box. "Chasing Equilibrium: Measuring the Intrinsic Solubility of Weak Acids and Bases". In: *Analytical Chemistry* 77.4 (Feb. 2005), pp. 983–990. DOI: 10.1021/ac048767n.

[54] A. J. Williams, C. M. Grulke, J. Edwards, A. D. McEachran, K. Mansouri, N. C. Baker, G. Patlewicz, I. Shah, J. F. Wambaugh, R. S. Judson, and A. M. Richard. "The CompTox Chemistry Dashboard: a community data resource for environmental chemistry". In: *Journal of Cheminformatics* 9.1 (Dec. 2017), p. 61. DOI: 10.1186/s13321-017-0247-6.

[55] J. L. Pascual-ahuir, E. Silla, and I. Tuñn. "GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface". In: *Journal of Computational Chemistry* 15.10 (Oct. 1994), pp. 1127–1138. DOI: 10.1002/jcc.540151009.

[56] M. Mantina, A. C. Chamberlin, R. Valero, C. J. Cramer, and D. G. Truhlar. "Consistent van der Waals Radii for the Whole Main Group". In: *The Journal of Physical Chemistry A* 113.19 (May 2009), pp. 5806–5812. DOI: 10.1021/jp8111556.

[57] M. Mons, I. Dimicoli, F. Piuzzi, B. Tardivel, and M. Elhanine. "Tautomerism of the DNA Base Guanine and Its Methylated Derivatives as Studied by Gas-Phase Infrared and Ultraviolet Spectroscopy". In: *The Journal of Physical Chemistry A* 106.20 (May 2002), pp. 5088–5094. DOI: 10.1021/jp0139742.

[58] D. W. Newton, S. Ratanamaneichatara, and W. J. Murray. "Dissociation, solubility and lipophilicity of azathioprine". In: *International Journal of Pharmaceutics* 11.3 (July 1982), pp. 209–213. DOI: 10.1016/0378-5173(82)90039-4.

[59] M. A. Makhyoun, R. A. Massoud, and S. M. Soliman. "Tautomerism and spectroscopic properties of the immunosuppressant azathioprine". In: *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 114 (Oct. 2013), pp. 394–403. DOI: 10.1016/j.saa.2013.05.041.

[60] N. L. Calvo, S. O. Simonetti, R. M. Maggio, and T. S. Kaufman. "Thermally induced solid-state transformation of cimetidine. A multispectroscopic/chemometrics determination of the kinetics of the process and structural elucidation of one of the products as a stable N3-enamino tautomer". In: *Analytica Chimica Acta* 875 (May 2015), pp. 22–32. DOI: 10.1016/j.aca.2015.02.033.

[61] L. A. Sheludyakova, t. l. E. V. Sobolev, A. V. Arbuznikov, E. B. Burgina, and L. I. Kozhevina. "Experimental and theoretical study of monosubstituted guanidines by vibrational spectroscopy Part 1.—Structure of cyanoguanidine". In: *Journal of the Chemical Society, Faraday Transactions* 93.7 (1997), pp. 1357–1360. DOI: 10.1039/a605916c.

[62] J. W. Black, G. J. Durant, J. C. Emmett, and C. R. Ganellin. "Sulphur-methylene isosterism in the developent of metiamide, a new histamine H2-receptor antagonist". In: *Nature* 248.5443 (Mar. 1974), pp. 65–67. DOI: 10.1038/248065a0.

[63] B. F. Spisso, M. A. G. de Araújo Júnior, M. A. Monteiro, A. M. B. Lima, M. U. Pereira, R. A. Luiz, and A. W. da Nóbrega. "A liquid chromatography–tandem mass spectrometry confirmatory assay for the simultaneous determination of several tetracyclines in milk considering keto–enol tautomerism and epimerization phenomena". In: *Analytica Chimica Acta* 656.1-2 (Dec. 2009), pp. 72–84. DOI: 10.1016/j.aca.2009.10.012.

[64] C. I. Miles and G. H. Schenk. "Fluorescence and phosphorescence of phenylethylamines and barbiturates. Analysis of amphetamine and barbiturate preparations". In: *Analytical Chemistry* 45.1 (Jan. 1973), pp. 130–136. DOI: 10.1021/ac60323a022.

[65] P. W. Kenny. *The Prediction of Tautomer Preference in Aqueous Solution*. 2019. DOI: 10.6084/m9.figshare.8966276.v1. URL: https://doi.org/10.6084/m9.figshare.8966276.v1.

[66] K. Takács-Novák, B. Noszál, I. Hermecz, G. Keresztúri, B. Podányi, and G. Szasz. "Protonation Equilibria of Quinolone Antibacterials". In: *Journal of Pharmaceutical Sciences* 79.11 (Nov. 1990), pp. 1023–1028. DOI: 10.1002/jps.2600791116.

[67] A. Felczak, U. Kalinowska-Lis, J. Kusz, and L. Checińska. "Zwitterionic versus neutral molecules of fluoroquinolones: crystal structure of danofloxacin dihydrate". In: *Acta Crystallographica Section C Structural Chemistry* 78.12 (Dec. 2022), pp. 722–729. DOI: 10.1107/S2053229622010300.

[68] V. Bhatt and R. Jee. "Micro-ionization acidity constants for tetracyclines from fluorescence measurements". In: *Analytica Chimica Acta* 167 (1985), pp. 233–240. DOI: 10.1016/S0003-2670(00)84425-6.

[69] J. Bordner, P. D. Hammen, and E. B. Whipple. "Deuterium isotope effects on carbon-13 NMR shifts and the tautomeric equilibrium in N-substituted pyridyl derivatives of piroxicam". In: *Journal of the American Chemical Society* 111.17 (Aug. 1989), pp. 6572–6578. DOI: 10.1021/ja00199a015.

[70] L. Guasch, M. L. Peach, and M. C. Nicklaus. "Tautomerism of Warfarin: Combined Chemoinformatics, Quantum Chemical, and NMR Investigation". In: *The Journal of Organic Chemistry* 80.20 (Oct. 2015), pp. 9900–9909. DOI: 10.1021/acs.joc.5b01370.

[71] Z. Zhou. *tfkan*. 2024. URL: https://github.com/ZPZhou-lab/tfkan.

[72] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. *Self-Normalizing Neural Networks*. 2017. arXiv: 1706.02515 [cs.LG]. URL: https://arxiv.org/abs/1706.02515.

[73] I. Loshchilov and F. Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG]. URL: https://arxiv.org/abs/1711.05101.