nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00856-0

# Molecular set representation learning

Received: 16 October 2023

Accepted: 21 May 2024

Published online: 5 July 2024

Check for updates

Maria Boulougouri, Pierre Vandergheynst & Daniel Probst

Computational representation of molecules can take many forms, including graphs, string encodings of graphs, binary vectors or learned embeddings in the form of real-valued vectors. These representations are then used in downstream classification and regression tasks using a wide range of machine learning models. However, existing models come with limitations, such as the requirement for clearly defined chemical bonds, which often do not represent the true underlying nature of a molecule. Here we propose a framework for molecular machine learning tasks based on set representation learning. We show that learning on sets of atom invariants alone reaches the performance of state-of-the-art graph-based models on the most-used chemical benchmark datasets and that introducing a set representation layer into graph neural networks can surpass the performance of established methods in the domains of chemistry, biology and material science. We introduce specialized set representation-based neural network architectures for reaction-yield and protein–ligand binding-affinity prediction. Overall, we show that the technique we denote molecular set representation learning is both an alternative and an extension to graph neural network architectures for machine learning tasks on molecules, molecule complexes and chemical reactions.

Chemical and biochemical structures and processes are of great interest to machine learning with graph neural networks (GNNs), as they are supposedly well represented by undirected and directed graphs, respectively[1]. Conversely, machine learning plays an ever-increasing role in the life sciences, where new methods have been adopted and adapted for a wide range of tasks such as the prediction of physico- and quantum-chemical properties in material science, the prediction of pharmacokinetic properties in drug development, the prediction of binding affinities between small-molecule ligands and proteins in drug discovery, or the prediction of chemical reaction yields in synthetic and process chemistry[2–6].

With the introduction of GNNs, graph representation learning for molecules attracted a growing interest, as these neural network architectures allowed representation learning directly on the molecular graphs rather than first extracting features or precalculating descriptors from a given chemical structure[7], with newly proposed methods continuously escalating the architectures' complexity or amount of precomputed or, less likely, experimentally measured molecular properties, such as three-dimensional coordinates or structural motifs, that are added to the graph[8–10]. However, even though the published literature often touts the benefits of introducing new features based on molecular composition and topology, it is generally ignored that these features are also implicitly encoded by lower-order representations. For example, the most-used text representation of molecules, simplified molecular-input line-entry system (SMILES), are a string encoding of the depth-first tree traversal of the molecular graph retaining all topological information and, in the case of isomeric SMILES, geometrical information pertaining to stereochemistry[11]. The same applies to molecular fingerprints, which often even explicitly encode topological features[12]. Meanwhile, the molecular graph introduces implicit constraints on the molecular three-dimensional geometry, as topology and atom types induce structure. Conversely, representing a molecule as a graph often encapsulates only an approximation of molecular topology and geometry, as only a subset of chemical bonds, the covalent bonds, can be represented in this static data structure. However, in addition to covalent bonds, molecules and substances used as drugs or materials also contain ionic and metallic bonds that are generally not well represented in molecular graphs. In addition,

Signal Processing Laboratory 2, Institute of Electrical and Micro Engineering, School of Engineering, EPFL, Lausanne, Switzerland.
e-mail: daniel.probst@epfl.ch

https://doi.org/10.1038/s42256-024-00856-0

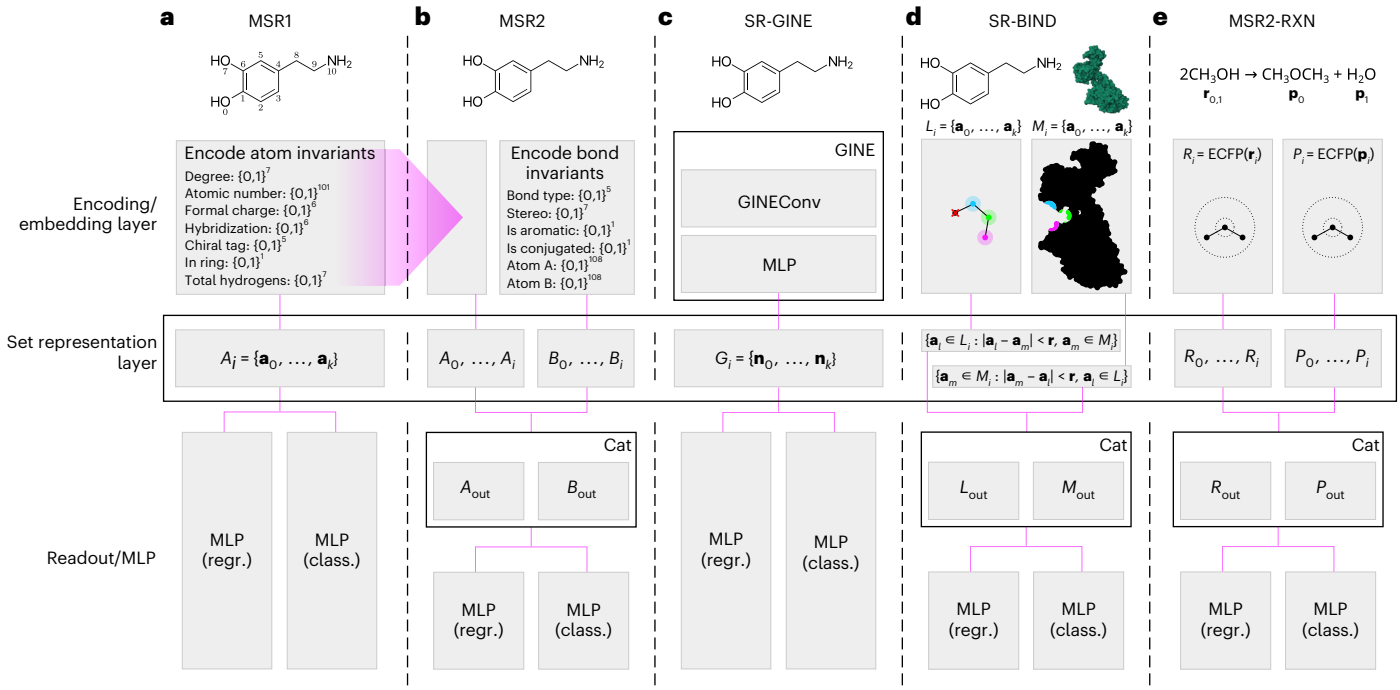


**Fig. 1 | Overview of set-based and set-enhanced models.** All implemented models consist of three parts: an encoding or embedding layer, a set representation layer and, finally, a readout/MLP layer. **a**, The simplest molecular set representation model MSR1 takes molecules as input and encodes each atom as 133-dimensional binary vectors $\mathbf{a}_0$ to $\mathbf{a}_k$, where $k$ is the number of atoms in the $i$th molecule, into a molecular set $A_i$. The resulting molecular sets with differing cardinalities are passed into a RepSet set representation layer and read out by a regression (regr.) or classification (class.) MLP. **b**, The dual molecular set representation model MSR2 encodes the atoms and bonds of molecules into two distinct sets $A_i$ and $B_i$ and passes them to two separate RepSet layers whose outputs $A_{out}$ and $B_{out}$ are concatenated (Cat) followed by either a regression or classification MLP. **c**, SR-GINE is a GIN model with a GINEConv layer enhanced with a set representation layer replacing global pooling. The node (atom) embeddings $\mathbf{n}_0$ to $\mathbf{n}_k$ are then passed to a RepSet layer as graph sets $G_i$ followed by an MLP regressor or classifier. **d**, SR-BIND follows the dual-set architecture of MSR2 by employing two parallel RepSet layers. Atoms $\mathbf{a}_i$ of the ligand $L_i$ are added to a set if they are within radius $r$ of any protein atom $\mathbf{a}_m$. Conversely, atoms $\mathbf{a}_m$ from protein $M_i$ are added to a second set only if they are within radius $r$ of any ligand atom $\mathbf{a}_i$. Both sets are passed to separate RepSet layers whose output is concatenated and passed to a regression or classification MLP. **e**, MSR2-RXN also follows the dual-set architecture of MSR2 by employing two parallel RepSet layers. All reactants $\mathbf{r}_j$ and products $\mathbf{p}_j$ that are part of the $i$th reaction in an input data set are encoded using ECFP with a radius of 3 and size of 2,048 into molecular sets $R_i$ and $P_i$. Both sets are passed to separate RepSet layers whose output is concatenated and passed to a regression or classification MLP.

dynamic intermolecular interactions such as hydrogen bonds and $\pi$-stacking are common occurrences in ligand–protein binding, which is of high interest in medicinal chemistry[13]. Furthermore, specific bonds are not well defined in conjugated systems such as aromatic rings, as electrons are delocalized over multiple atoms and bonds.

Given this often somewhat fuzzy notion of bonds in molecules, we hypothesize that representing a molecule as a set—formally a multiset, as the vectors representing the atoms can be identical for two or more atoms—of atoms rather than a graph may capture the true nature of molecules better than explicit graph representations while preserving implicit information about molecular structure. In this set-based approach, each atom is represented as a vector of one-hot encoded atom invariants as defined in ref. 12 (see details in 'Choice of atom and bond invariants' in Methods). While this representation may encode the local topology of a molecular graph through these invariants, for example, the degree of an atom, it does not encode any explicit connectivity of the molecular graph. In the context of set representation learning, a molecule is therefore defined as a set of $k$-dimensional vectors $\mathbf{a}$ where the set's cardinality is the number of non-hydrogen atoms in the molecule. However, this set-based representation introduces two properties that are not supported by typical neural network-based machine learning architectures: (1) the cardinality of the molecular sets differs depending on the number of non-hydrogen atoms in a molecule, and (2) the molecular sets are unordered. In addition, the architecture must support multiset input, as the number of identical vector representations of atoms matters in the context of molecular sets. Given these requirements, specifically the need for permutation invariance, it is insufficient to simply pad the sets. Therefore, performing machine learning tasks on such molecular sets requires a neural network architecture capable of permutation invariant representation of variable-sized sets. Over the past five years, multiple architectures, including DeepSets, Set-Transformer or RepSet, have become available that fulfil these requirements[14–16].

Here we introduce several of what we denote molecular set representation learning architectures based on the scheme described above and in Fig. 1, and evaluate them against widely used GNN methods on a wide array of tasks, including physico- and quantum-chemical property prediction for material science, pharmacokinetic property prediction for drug design, protein–ligand-affinity prediction for drug discovery and biochemistry, and reaction-yield prediction for synthetic and process chemistry. We show that using the concept of molecular set representation learning and combining it with GNN architectures, we can not only simplify current approaches but also improve on commonly used GNN implementations such as directed message-passing neural network (D-MPNN), graph attention network (GAT) or DimeNet[17–19]. Furthermore, we uncover that compared with more modern benchmark datasets, extensively used older benchmark datasets may not be well suited to evaluate the advantages of GNNs, as they perform worse than the most simple of our models. Finally, we introduce an easy-to-use, extensible collection of molecular set representation architectures ready to be used in various fields, including materials science, drug discovery and development, biochemistry, and synthetic and process chemistry.

**Table 1 | Benchmarking the implemented molecular set representation architectures against GNN baselines on the MoleculeNet datasets**

| Dataset | Metric | MolFormer | MGCN | SchNet | GCN | GIN | D-MPNN | MSR1 | MSR2 | GINE | SR-GINE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HIV | AUROC | 81.3 | 73.8±1.6 | 70.2±3.4 | 71.6±4.0 | 75.3±1.9 | 75.0±2.1 | 72.3±2.2 | **75.6±1.7** | 74.3±1.9 | 71.0±1.0 |
| BACE | AUROC | 86.6 | 73.4±3.0 | 76.6±1.1 | 71.6±2.0 | 70.1±5.4 | **85.3±5.3** | 75.5±1.9 | 76.6±2.0 | 63.1±1.6 | 77.7±1.1 |
| BBBP | AUROC | 91.5 | **85.0±6.4** | 84.8±2.2 | 71.8±0.9 | 65.6±0.5 | 71.2±3.8 | 71.4±0.0 | 70.7±1.7 | 64.0±1.6 | 69.3±0.8 |
| Tox21 | AUROC | 84.5 | 70.7±1.6 | **77.2±2.3** | 70.9±2.6 | 74.0±0.8 | 68.9±1.3 | 72.1±5.0 | 72.7±5.1 | 69.5±6.8 | 68.8±7.1 |
| SIDER | AUROC | 68.9 | 55.2±1.8 | 53.9±3.7 | 53.6±3.2 | 57.3±1.6 | **63.2±2.3** | 61.4±7.3 | 61.2±7.2 | 55.6±8.1 | 58.0±6.6 |
| ClinTox | AUROC | 94.6 | 63.4±4.2 | 71.5±3.7 | 62.5±2.8 | 58.0±4.4 | **90.5±5.3** | 86.6±1.2 | 80.1±2.6 | 83.9±1.9 | 87.0±2.1 |
| ESOL | RMSE | 0.33 | 1.27±0.15 | 1.05±0.06 | 1.43±0.05 | 1.45±0.02 | 0.98±0.26 | **0.59±0.03** | 0.59±0.05 | 0.98±0.10 | 0.73±0.17 |
| FreeSolv | RMSE | 0.22 | 3.35±0.01 | 3.22±0.76 | 2.87±0.14 | 2.76±0.18 | 2.18±0.91 | **1.94±0.24** | 2.07±0.05 | 2.92±0.31 | 2.52±0.45 |
| Lipo | RMSE | 0.54 | 1.11±0.04 | 0.91±0.10 | 0.85±0.08 | 0.85±0.07 | **0.65±0.05** | 0.85±0.03 | 0.82±0.01 | 0.81±0.01 | 0.76±0.02 |
| QM7 | MAE | – | 77.6±4.7 | **74.2±6.0** | 122.9±2.2 | 124.8±0.7 | 105.8±13.2 | 85.9±13.1 | 89.6±2.9 | 135.5±19.0 | 90.8±5.0 |
| QM8 | MAE | 0.011 | 0.022±0.002 | 0.020±0.002 | 0.037±0.001 | 0.037±0.001 | **0.014±0.002** | 0.023±0.010 | 0.022±0.011 | 0.025±0.012 | 0.025±0.010 |

For each dataset, the best value after MolFormer is highlighted in bold; the better performance between D-MPNN and MSR1, and between GINE and SR-GINE, respectively, is underlined. Training and testing were run three times; mean±s.d. are reported. The model reported as GIN by ref. 20 also includes a GINEConv layer. Compared with the set-layer comparisons, the RepSet-specific parameters of SR-GINE were further tuned. The metrics for the benchmarks are area under the receiver operating characteristic (AUROC) curve, root mean square error (RMSE) and mean absolute error (MAE). HIV, human immunodeficiency virus.

## Results and discussion

We introduce multiple implementations of molecular set representation architectures, including the single-set atom-based MSR1, the dual-set atom and bond-based MSR2, and the graph invariant network-based set representation-graph isomorphism network with edge (SR-GINE) that replaces the pooling function with a set representation layer. For input into MSR1, a molecule is encoded as a set of one-hot encoded vectors, where each vector encodes the atom invariants of a single atom in the molecule, similar to extended-connectivity fingerprint (ECFP) with radius zero, containing no explicit information about molecular topology. Following refs. 12,17, we include atom invariants described in 'Choice of atom and bond invariants' in Methods. MRS2 expands on this concept and implements a neural network architecture with two parallel set representation layers, where the first takes the same input as the single layer in MSR1. In contrast, the second takes a set of vectors encoding bond invariants as input. We include bond invariants described in 'Choice of atom and bond invariants' in Methods. Therefore, with MSR2, we introduce more information about molecular topology while avoiding explicit definitions of topology beyond bonded atom pairs. In addition to the purely set-representation architectures, we introduce SR-GINE, a graph invariant network with edge attributes (GINE) and a set RepSet representation pooling layer instead of global mean pooling in the vanilla GINE implementation. In all our architectures, the output of the set representation layer is read by a multilayer perceptron (MLP) with a single hidden layer for regression or classification. In the case of dual-set architectures, such as MSR2, the two sets' outputs are first concatenated. Finally, we chose the combination of RepSet and GINE based on their respective reported performances and the inert interpretability of RepSet[7,16]. Further reasoning and data substantiating our choice can be found in 'Choice of set representation and GNN layers' in Methods.

To benchmark our proposed architectures, we rely on well-known and recently published datasets to compare our method with widely used graph-based methods as a baseline. Initially, we focused on evaluating our methods on datasets commonly known as the MoleculeNet benchmark. We selected the relatively small dataset blood–brain barrier-penetrating molecules (BBBP) ($n$ = 2,039), with the task of classifying small molecules on whether they penetrate the blood–brain barrier, to tune the hyperparameters, namely, the number of hidden sets, the number of elements in the hidden sets, the number of epochs, as well as the number of hidden channels in the MLPs of all our models. Unless otherwise indicated, the established hyperparameters were used for all the benchmarks discussed in this study. We then compared our model with the performance achieved by multiscale graph convolutional network (MGCN), SchNet (quantum-chemical deep tensor neural network), graph convolutional network (GCN), graph isomorphism network (GIN) and D-MPNN, as reported in ref. 20. In addition, we compare our approach with the current state of the art outside GNN-based approaches, namely, MolFormer, a chemical large language model trained on approximately 100 million molecules extracted from the ZINC and PubChem databases[21]. As a control, we benchmark an implementation of GINE with the standard global mean-pooling layer and the same one-hot encoded vectors as atom and bond attributes that were used for MSR1 and MSR2[7]. For all datasets, Murcko scaffold splits in accordance with refs. 17,20 were used.

As shown in Table 1, MSR1, the simplest of our models, shows a performance close to existing GNN approaches, namely, GIN and D-MPNN, without any explicit topological information about the molecular graph. Indeed, it performed better than D-MPNN in 5 out of 11 benchmark datasets and better than GIN in 8 out of 11. This may suggest that up to now, too much value has been assigned to representing a molecule as a graph rather than a loose set of atoms. Alternatively, these results may be due to the nature of the benchmark datasets or the molecules contained therein, meaning that they are potentially not suited for evaluating GNN architectures. Furthermore, MSR2 did not improve on the performance of MSR1 as expected but performed generally worse. Finally, replacing the global mean-pooling layer of GINE with RepSet improved its performance in 8 out of 11 benchmarks. Overall, these results are promising in light of our hypothesis, which stipulates that a more relaxed definition of molecules than the one provided by graph encoding may be beneficial. However, as discussed previously, the performance of our simplistic models, especially MSR1, may be due to limitations of the datasets. Hence, we benchmarked our architectures on two recent and well-received datasets, the results of which are discussed in 'Physico- and quantum-chemical property prediction' and 'Pharmacokinetic property prediction'. In addition to these more generalized architectures, we introduce two architectures based on the dual-set approach of MSR2 tailored towards binding-affinity prediction in protein–ligand complexes and reaction-yield prediction in 'Binding-affinity prediction' and 'Reaction-yield prediction', respectively.

### Physico- and quantum-chemical property prediction

The prediction of physicochemical properties is a common task in machine learning for chemistry, as the elucidation of these

**Table 2 | Benchmarking the implemented molecular set representation architectures against reported baselines on the OCELOT chromophore data containing chemically diverse π-conjugated molecules**

| Property | MolFormer | ECFP2+ | MPNN | MPNN+ | MSR1 | MSR2 | GINE | SR-GINE |
|---|---|---|---|---|---|---|---|---|
| HOMO | 0.477±0.012 | 0.354±0.012 | 0.796±0.446 | 0.330±0.028 | 0.348±0.011 | 0.342±0.001 | 0.335±0.009 | **0.328±0.007** |
| LUMO | 0.267±0.006 | 0.297±0.004 | 0.291±0.044 | 0.289±0.028 | 0.324±0.004 | 0.298±0.004 | 0.285±0.006 | **0.269±0.007** |
| H–L | 0.674±0.007 | 0.578±0.011 | 1.264±0.696 | **0.548±0.029** | 0.603±0.017 | 0.583±0.001 | 0.581±0.032 | 0.561±0.018 |
| VIE | 0.350±0.005 | 0.219±0.001 | 0.202±0.043 | 0.191±0.024 | 0.247±0.004 | 0.230±0.005 | 0.199±0.005 | **0.183±0.003** |
| AIE | 0.351±0.006 | 0.207±0.003 | 0.176±0.015 | 0.173±0.006 | 0.240±0.003 | 0.220±0.004 | 0.188±0.004 | **0.172±0.004** |
| CR1 | 0.052±0.001 | 0.063±0.001 | **0.054±0.001** | 0.055±0.002 | 0.061±0.001 | 0.059±0.001 | 0.059±0.002 | 0.057±0.000 |
| CR2 | 0.051±0.000 | 0.059±0.001 | 0.061±0.001 | **0.053±0.001** | 0.061±0.000 | 0.060±0.001 | 0.057±0.001 | 0.056±0.001 |
| HR | 0.098±0.001 | 0.110±0.002 | 0.126±0.022 | 0.133±0.019 | 0.117±0.001 | 0.114±0.002 | 0.110±0.002 | **0.109±0.002** |
| VEA | 0.133±0.000 | 0.186±0.002 | 0.193±0.052 | 0.157±0.018 | 0.252±0.005 | 0.212±0.003 | 0.164±0.005 | **0.154±0.005** |
| AEA | 0.117±0.001 | 0.176±0.002 | 0.160±0.027 | 0.154±0.027 | 0.244±0.006 | 0.203±0.004 | 0.154±0.004 | **0.141±0.004** |
| AR1 | 0.049±0.000 | 0.062±0.002 | 0.057±0.002 | **0.051±0.001** | 0.060±0.002 | 0.059±0.001 | 0.057±0.001 | 0.054±0.001 |
| AR2 | 0.045±0.001 | 0.051±0.001 | **0.048±0.002** | 0.052±0.001 | 0.054±0.001 | 0.053±0.001 | 0.051±0.001 | 0.050±0.001 |
| ER | 0.090±0.002 | 0.101±0.002 | **0.093±0.002** | 0.098±0.006 | 0.110±0.002 | 0.107±0.002 | 0.102±0.002 | 0.099±0.002 |
| S0S1 | 0.292±0.007 | 0.282±0.003 | 0.252±0.017 | **0.249±0.013** | 0.356±0.003 | 0.307±0.004 | 0.268±0.006 | 0.256±0.007 |
| S0T1 | 0.162±0.001 | 0.194±0.003 | **0.148±0.012** | 0.150±0.028 | 0.297±0.007 | 0.235±0.002 | 0.175±0.004 | 0.160±0.006 |
| Average | 0.213±0.186 | 0.196±0.142 | 0.261±0.333 | 0.178±0.134 | 0.225±0.153 | 0.205±0.143 | 0.186±0.141 | **0.177±0.136** |

For each dataset, the best value is highlighted in bold; the better performance between MPNN and MSR1, and between GINE and SR-GINE, respectively, is underlined. MolFormer results are not formatted to facilitate a comparison between the graph- and set-based methods. Training and testing were run five times (three times for the finetuned MolFormer); mean±s.d. are reported. The metric used is MAE. In addition to the HOMO and LUMO values, the predicted properties are the HOMO–LUMO energy gap (H–L), the vertical and adiabatic ionization energies (VIE and AIE), cation and anion relaxation energies (CR1,2 and AR1,2), electron and hole reorganization energies (ER and HR), vertical and adiabatic electron affinities (VEA and AEA), and the lowest-lying singlet and triplet excitation energies (S0S1 and S0T1).

properties through either laboratory high-throughput screening or simulation-based computation is expensive and time-consuming at best and intractable at worst[22,23]. Hence, multiple approaches have been proposed to enable data-driven approximation of properties such as frontier molecular orbital energies, the molecular dipole moment, rotational constants or ionization potentials. During our exploratory study of our proposed architectures on the quantum-chemical benchmark datasets QM7 and QM8 (Table 1), we found that our simplest model (MSR1) showed better performance than both GIN(E) and D-MPNN on QM7 and better performance than GIN(E) on QM8. In this section, we further investigate the performance of our models on additional physico- and quantum-chemical prediction tasks using the OCELOT chromophore dataset[24].

The OCELOT chromophore dataset contains chemically diverse π-conjugated molecules, meaning that our set-based architectures should, according to our hypothesis, perform better than graph-based architectures as they put less emphasis on specific bonds[24]. For this dataset, we changed our hyperparameters to reduce the size of all our models to less than 100,000 parameters, as the relatively high number of samples and tasks, combined with the 5-fold cross-validation, requires non-trivial computational resources. The original study used the dataset to train a hierarchy of models that follow increasingly complex architectures, peaking with an MPNN for quantum chemistry[25]. In addition, the output of the MPNN was concatenated with precomputed molecular descriptors before the feedforward neural network. We denote this model MPNN+ in Table 2. Furthermore, a method based on a fingerprint-descriptor combination, which we donate ECFP2+, showed exceptional performance in the original study. ECFP2+ is a feedforward neural network that takes ECFP fingerprints with radius $r = 2$ concatenated with precomputed molecular descriptors as an input.

Our most straightforward model, MSR1, performs as well as MPNN over all predicted properties (paired $t$-test, $P = 0.529$), as MPNN performs remarkably poorly on the highest occupied molecular orbital (HOMO) and HOMO−lowest unoccupied molecular orbital (LUMO) (H–L) tasks. However, both set-based models, MSR1 and MSR2, perform significantly worse than MPNN+ (paired $t$-test $P < 0.001$ and $P < 0.002$, respectively). These results suggest that adding explicit bonds and introducing message-passing does not significantly improve the predictive accuracy on the OCELOT dataset, and to improve performance, additional molecular precomputed descriptors are needed; however, in practice, the choice of model may be influenced by the property of interest given the poor performance of MPNN on specific tasks, namely, HOMO and H–L. Over all properties, the set-enhanced model SR-GINE performs significantly better than GINE (paired $t$-test, $P < 0.0001$), yet not significantly different from the MPNNs (paired $t$-test, $P < 0.138$), although without the significant outliers observed when predicting HOMO and the H–L gap with MPNN, or the additional precomputed RDKit two-dimensional descriptors required by MPNN+ (paired $t$-test, $P < 0.409$).

To compare our set-based approach to a current state-of-the-art chemical large language model, we finetuned the publicly available pretrained variant of MolFormer on OCELOT ('Finetuning MolFormer' in Methods). Interestingly, the finetuned MolFormer performed below expectations compared with the MoleculeNet benchmark. A reason for the comparatively lower performance could be a lack of relevant samples in the original training set of the publicly available variant of MolFormer[26,27].

Together with the results on QM7 and QM8 (Table 1), the results of the OCELOT chromophore benchmark suggest that set representation learning can perform as well as graph representation learning on physico- and quantum-chemical prediction tasks. Furthermore, combining graph and set representation learning, as is done with SR-GINE, shows synergistic effects that result in overall enhanced performance. Specifically, the significant performance increase of SR-GINE over GINE indicates that a set representation layer in lieu of a global pooling function can improve existing GNN architectures.

## Pharmacokinetic property prediction

A molecule's pharmacokinetic properties play an essential role in designing and developing new therapeutics[28]. However, experimental

**Table 3 | Benchmarking the implemented molecular set representation architectures against reported baselines on the data from Biogen ADME in vitro assays**

| Endpoint | MolFormer | RF | LightGBM | D-MPNN | D-MPNN+ | MSR1 | MSR2 | GINE | SR-GINE |
|---|---|---|---|---|---|---|---|---|---|
| HLM | 0.71±0.00 | 0.62 | 0.67 | 0.65 | **0.68** | 0.58±0.00 | 0.61±0.02 | 0.59±0.01 | **0.68±0.01** |
| MDR1–MDCK efflux ratio | 0.82±0.00 | 0.73 | 0.76 | 0.72 | 0.78 | 0.68±0.02 | 0.70±0.02 | 0.67±0.00 | **0.80±0.01** |
| Solubility | 0.60±0.01 | 0.58 | 0.56 | **0.64** | 0.59 | 0.47±0.04 | 0.52±0.04 | 0.55±0.02 | 0.63±0.02 |
| RLM | 0.73±0.00 | 0.66 | 0.68 | 0.72 | **0.74** | 0.57±0.01 | 0.57±0.02 | 0.61±0.02 | 0.70±0.01 |
| hPPB | 0.77±0.00 | 0.76 | 0.80 | 0.74 | **0.77** | 0.68±0.01 | 0.72±0.01 | 0.68±0.02 | **0.77±0.01** |
| rPPB | 0.74±0.00 | 0.67 | 0.73 | 0.70 | 0.70 | 0.55±0.02 | 0.61±0.02 | 0.63±0.02 | **0.73±0.02** |
| Average | 0.729±0.005 | 0.670±0.061 | 0.700±0.077 | 0.695±0.037 | 0.710±0.064 | 0.588±0.08 | 0.622±0.08 | 0.622±0.05 | **0.718±0.06** |

For each dataset, the best value is highlighted in bold; the better performance between D-MPNN and MSR1, and between GINE and SR-GINE, respectively, is underlined. MolFormer results are not formatted to facilitate a comparison between the graph- and set-based methods. Training and testing were run three times; mean±s.d. are reported. RF and LightGBM use 1,024-bit FCFP4 fingerprints concatenated with 316 RDKit descriptors as input. The metric used is Pearson's correlation coefficient as suggested by ref. 31.

pharmacokinetic data from in vitro and vivo experiments is scarce, as it is expensive to generate, making its approximation a priority for machine learning research[29,30]. To evaluate our models' performance on pharmacokinetic tasks, we use the dataset recently published by ref. 31. The dataset contains 3,521 commercially available compounds that were tested against the following endpoints in Biogen absorption, distribution, metabolism and excretion (ADME) in vitro assays: human liver microsomal stability (HLM, clearance in ml $min^{-1}$ $kg^{-1}$), MDR1–MDCK efflux ratio (permeability with Madin Darby–canine kidney cells transfected with MDR1), solubility at a pH of 6.8 (μg $ml^{-1}$), rat liver microsomal stability (RLM, clearance in ml $min^{-1}$ $kg^{-1}$), human plasma protein binding (hPPB, percent unbound) and rat plasma protein binding (rPPB, percent unbound). Reference 31 provides baselines trained on the dataset for random forests (RF), gradient boosting (LightGBM), a hyperparameter-tuned D-MPNN and a hyperparameter-tuned D-MPNN plus precomputed RDKit two-dimensional descriptors (D-MPNN+)[17]. Values for additional models can be found in the original publication. As input for RF and LightGBM, ref. 31 concatenated 1,024-bit binary functional connectivity fingerprint with radius 4 (FCFP4) concatenated with 316 two-dimensional descriptors that were precomputed using the RDKit package[12]. In addition, we finetuned the publicly available pretrained variant of MolFormer on the Biogen ADME dataset ('Finetuning MolFormer' in Methods). The benchmark results from the original publication are compared to ours and those of MolFormer in Table 3.

Over all pharmacokinetic endpoints, SR-GINE performs significantly better than the control GINE (paired $t$-test, $P = 1.98 \times 10^{-5}$), which uses the standard mean pooling instead of a set layer. Furthermore, the SR-GINE model performs better than the hyperparameter-tuned D-MPNN and D-MPNN+ models without relying on hyperparameter tuning or additional precomputed descriptors (RDKit two-dimensional descriptors). While MSR1 and MSR2 perform worse than D-MPNN(+) and SR-GINE, they do not perform significantly differently to GINE (paired $t$-test, $P = 0.096$ and $P = 1.0$, respectively). In other words, they perform as well as an out-of-the-box state-of-the-art GNN. The results of MSR1 compared with D-MPNN are interesting, as they contradict drug discovery and design-related findings from benchmarks on the MoleculeNet datasets, where MSR1 often compared favourably to D-MPNN, a prominent example being solubility prediction (Table 1). As the benchmarks found in MoleculeNet remain the most used and cited when benchmarking new GNN architectures, our results may show a need to adopt a different practice or update MoleculeNet with more recent datasets when assessing possible advantages of graph representation-based approaches. In addition, the difference in performance between graph- and set-based methods and that of the chemical large language model (MolFormer) is not as stark on the Biogen ADME dataset as it was for the MoleculeNet benchmarks. Indeed, MolFormer does not perform significantly better than SR-GINE (paired $t$-test, $P = 0.158$).

Finally, the renewed significant improvement of SR-GINE over GINE strengthens the case for improving predictive performance by enhancing GNN architectures with a set representation layer.

## Binding-affinity prediction

Binding affinity is an important metric in biology and medicinal chemistry that measures the strength of a reversible association between biological macromolecules, such as proteins or DNA, and small-molecule ligands, such as drugs. It is, therefore, a central concept of rational drug design, where the potential efficacy of a drug is measured by its binding affinity to a known biological target implicated in a pathology the drug should treat[32]. Traditionally, simulation-based molecular docking techniques, such as scoring functions, overfit to use cases that adhere to rigid modelling assumptions, do not fully represent protein flexibility, and do not directly account for solvent effects[33,34]. Non-parametric machine learning has been proposed as an alternative to infer complex binding effects that are difficult to explicitly represent directly from experimental data[35]. Therefore, a wide array of neural network architectures combined with a multitude of scoring functions have been proposed[36–40].

However, representation learning methods—foremost molecular graph representation learning—have yet to reach the performance established by physics-informed methods such as PAMNet[41], or RF and gradient-boosting-based approaches that make use of engineered features and precomputed molecular descriptors such as the ECFP-inspired extended connectivity interaction features (ECIF)[42].

For our model used for binding-affinity prediction, we borrowed the architecture and all hyperparameters from MSR2. The encoding of the ligand–protein complex consists of creating a multiset of atoms for the ligand and one for the protein, respectively. The set $L$ representing the ligand is constructed by iterating over the atoms of the ligand, and adding those that are within a radius of $r$ from any atom in the protein to the set. The set $M$ representing the protein is constructed in the same way, however, with the roles of the ligand and protein reversed. The optimal value of $r = 5.5$ was determined by treating it as a hyperparameter and evaluating its effect on the validation set and supports the findings by ref. 42. All other hyperparameters were kept from the original MSR2 architecture.

We evaluated our method based on PDBbind splits and metrics of basic GNN methods architectures reported by ref. 39. SR-BIND compares well to all graph drug–target affinity (GraphDTA) and GNN-based methods[40] (Table 4). This result suggests that, compared with the distance between specific protein and ligand atoms, the molecular topology of the ligand plays a relatively minor role. It must noted that the various GraphDTA methods do not use geometric information but rely on the more readily available amino acid sequence information for representing the protein.

**Table 4 | Performance of our set-based method SR-BIND compared against GNN-based methods**

| Model | | RMSE | MAE | *R* |
|---|---|---|---|---|
| GraphDTA methods | GCN | 1.735±0.034 | 1.343±0.037 | 0.613±0.016 |
| | GAT | 1.765±0.026 | 1.354±0.033 | 0.601±0.016 |
| | GIN | 1.640±0.044 | 1.261±0.044 | 0.667±0.018 |
| | GAT-GCN | 1.562±0.022 | 1.191±0.016 | 0.697±0.008 |
| GNN-based methods | SGCN | 1.583±0.033 | 1.250±0.036 | 0.686±0.015 |
| | GNN-DTI | 1.492±0.025 | 1.192±0.032 | 0.736±0.021 |
| | D-MPNN | 1.493±0.016 | 1.188±0.009 | 0.729±0.006 |
| | MAT | 1.457±0.037 | 1.154±0.037 | 0.747±0.013 |
| | DimeNet | 1.453±0.027 | 1.138±0.026 | 0.752±0.010 |
| | CMPNN | 1.408±0.028 | **1.117±0.031** | 0.765±0.009 |
| Set-based methods | SR-BIND | **1.383±0.049** | 1.122±0.041 | **0.780±0.012** |

SR-BIND compares favourably against all methods, due to its flexibility that allows for easy integration of geometrical features. Training and testing were run four times; mean±s.d. are reported. The value representing the best performance for each metric is highlighted in bold. The acronyms SGCN, DTI, MAT and CMPNN stand for spatial graph convolutional network, drug–target interaction, Molecule Attention Transformer and Communicative Message Passing Neural Network, respectively.

**Table 5 | Performance of MSR2-RXN compared with the state of the art in reaction-yield prediction on experimentally determined yields of Buchwald–Hartwig reactions through HTEs and extracted from an ELN**

| Dataset | Subset/spilt | DFT | Yield-BERT | Yield-BERT (aug.) | DRFP | YieldGNN | MSR2-RXN |
|---|---|---|---|---|---|---|---|
| Buchwald–Hartwig (HTE) | Rand 70/30 | 0.92 | 0.95±0.005 | **0.97±0.003** | 0.95±0.005 | 0.96±0.005 | 0.94±0.005 |
| | Rand 50/50 | 0.9 | 0.92±0.01 | **0.95±0.01** | 0.93±0.01 | – | 0.93±0.01 |
| | Rand 30/70 | 0.85 | 0.88±0.01 | **0.92±0.01** | 0.89±0.01 | – | 0.90±0.01 |
| | Rand 20/80 | 0.81 | 0.86±0.01 | **0.89±0.01** | 0.87±0.01 | – | 0.87±0.01 |
| | Rand 10/90 | 0.77 | 0.79±0.02 | **0.81±0.02** | **0.81±0.01** | – | 0.80±0.02 |
| | Rand 5/95 | 0.68 | 0.61±0.04 | **0.74±0.03** | 0.73±0.02 | – | 0.69±0.03 |
| | Rand 2.5/97.5 | 0.59 | 0.45±0.05 | 0.61±0.04 | **0.62±0.04** | – | 0.57±0.05 |
| | Test 1 | 0.8 | **0.84±0.01** | 0.80±0.01 | 0.81±0.01 | – | 0.83±0.03 |
| | Test 2 | 0.77 | 0.84±0.03 | **0.88±0.02** | 0.83±0.003 | – | 0.83±0.01 |
| | Test 3 | 0.64 | **0.75±0.04** | 0.56±0.08 | 0.71±0.001 | – | 0.69±0.04 |
| | Test 4 | **0.54** | 0.49±0.05 | 0.43±0.04 | 0.49±0.004 | – | 0.51±0.04 |
| | Average 1–4 | 0.69 | **0.73** | 0.58±0.33 | 0.71±0.16 | – | 0.72±0.15 |
| Buchwald–Hartwig (ELN) | – | – | −0.006±0.11 (0.25±0.01) | – | **0.20±0.05 (0.21±0.01)** | 0.05±0.07 (0.23±0.01) | 0.13±0.08 (0.21±0.12) |

Training and testing were run ten times (four times for tests 1–4); mean±s.d. are reported. The metric used is $R^2$ and for Buchwald–Hartwig (ELN); the MAE is in parentheses. For each subset/split combination, the value representing the best performance is highlighted in bold.

## Reaction-yield prediction

Predicting outcomes of chemical reactions, such as their yield based on data gathered in high-throughput screening, is an important task in machine learning for chemistry. Previously, we were able to show that relatively simple fingerprint-based gradient-boosting models can perform at least as well as computationally expensive density functional theory (DFT) and transformer-based methods[43]. In cheminformatics, chemical reactions are often defined as two sets of molecules, reactants and products, where the reactants are fully or partially transformed into the products during the reaction process. This set-based definition of chemical reactions hints at a potential use for set representation learning. Again, we focused on the most straightforward implementation by creating a dual-set neural network, where the inputs are binary ECFP (with $r = 3$ and including stereochemistry) vectors of reactant and product molecules, respectively.

We evaluated this architecture, denoted MSR2-RXN, against the state of the art using a high-throughput dataset of Buchwald–Hartwig cross-coupling reactions with the task of predicting reaction yields, that is, the percentage of input material (reactants) that is transformed to output materials (product). The baseline models include our previously introduced gradient-boosting and fingerprint-based method (DRFP), a DFT-based random-forest model (DFT), as well as the transformer-based models Yield-bidirectional encoder representations from transformers (BERT) and its augmented variant (Yield-BERT (aug.))[43]. MSR2-RXN performs similarly to the state-of-the-art methods on the ablation study (Rand $x/y$) using random splits and on the out-of-distribution splits (test $n$), as shown in Table 5. Indeed, MSR2-RXN does not perform significantly different compared with Yield-BERT, Yield-BERT (aug.) and DRFP (paired $t$-test, $P = 0.283$, $P = 1.000$ and $P = 0.307$, respectively) and performs significantly better than the DFT-based method (paired $t$-test, $P = 0.008$). In addition, while it also performs below a usable threshold on the Buchwald–Hartwig electronic laboratory notebook (ELN) dataset, it does perform better than both the BERT- and GNN-based models, coming closer in performance to DRFP.

The results illustrate the power and flexibility of molecular set representation learning. While the other methods rely on DFT calculations, pretrained large language models or a custom molecular fingerprint in the case of DRFP, MSR2-RXN uses the known simple and well-established ECFP embedding to represent the molecules

participating in the reaction. Given the development of advanced and specialized alternatives to ECFP[44], introducing our set-based architecture provides a new avenue for improving machine learning for chemical reactions using diverse molecular representations.

## Conclusion

With this initial foray into molecular set representation methodology, we were able to show competitive results of the technique across a wide range of use cases, including the prediction of quantum-chemical properties, pharmacokinetic properties, binding affinities and reaction yields with minimal hyperparameter adjustments and consistently straightforward architectures. Our most straightforward model, MSR1, which is essentially a set of ECFP fingerprints with a radius of zero, performs as well or better than D-MPNN on 5 out of 11 and better than GIN(E) on 9 out of 11 of the tested MoleculeNet benchmark datasets. With the poorer performance of the purely set representation-based methods MSR1 and MSR2 compared with SR-GINE, GINE and (D-)MPNN on more recent datasets, we have shown that the commonly used benchmarks provided by MoleculeNet and other cheminformatics and machine learning libraries may not be well suited to benchmark GNN architectures, as including explicit molecular topology does not seem to provide an advantage. Given the past and current reliance on these datasets of researchers when benchmarking new molecular deep learning architectures, we conclude and suggest that future architectures should be evaluated on other datasets, including those made available by refs. 24,31.

Furthermore, we showed that introducing a set representation layer in place of a global pooling function in a GNN (specifically a GINE) improves its performance in virtually all benchmarks and, in more recently released datasets, performs equal to or better than MPNN and D-MPNN without the need to introduce additional precomputed molecular descriptors. This insight may be used to extend and improve the performance of all currently used GNN-based molecular representation approaches. In addition, we introduced a set-based model for protein–ligand binding-affinity prediction, which allows for the introduction of implicit geometric information through a radius near-neighbour search between the protein and the bound ligand, allowing the model to perform better than existing graph-based approaches, which often cannot integrate such information easily. Finally, our conceptually naive set-based reaction-yield prediction model more than doubles the performance ($R^2$) of YieldGNN on the ELN-extracted dataset of Buchwald–Hartwig cross-coupling reactions and matches the performance of established methods on the high-throughput experiment (HTE) data.

In addition, an initial comparison between SR-GINE and MolFormer, a large language model pretrained on approximately 100 million compounds, shows no significant differences in performance on recently released benchmark datasets.

Overall, we present results that introduce and back the value of considering molecular set representation learning as an additional important branch of machine learning in computational chemistry and cheminformatics, as it provides a relaxation of the explicit molecular graph topology used in graph representation learning, which we showed to improve and extend capabilities across a wide range of tasks. The results of our set-based and set-enhanced methods on the OCELOT chromophore and PDBbind datasets also support our initial hypothesis that a more relaxed definition of molecular topologies enables the neural network to learn a more meaningful representation of molecules involving conjugated or transient bonds.

## Methods

### Study design

Initially, we evaluated our general approach against the well-known MoleculeNet datasets, excluding datasets that would require high amount of training time due to their size or number of tasks (ToxCast, MUV, PCBA). QM9 is used to evaluate the scalability of SR-GINE. Furthermore, we excluded PDBbind from our initial evaluation as our base models do not support protein–ligand complexes.

After the initial experiments, we chose more modern and specific benchmark datasets, namely, the OCELOT chromophore dataset and a set of compounds evaluated against Biogen ADME in vitro assays. Furthermore, we introduced two additional set-based models to handle protein–ligand complexes from the PDBbind dataset and reactions from the Buchwald–Hartwig HTE and ELN datasets.

### Choice of atom and bond invariants

The concept of atom and bond invariants has been borrowed from the Daylight atom invariants[11] and allows us to represent each atom and bond in the molecule as a one-hot encoded vector.

In all models, the atom and bond invariants were chosen based on the choices by refs. 12,17. As shown in Fig. 1a, the atom invariants are (1) the degree (total number of bonds) of the atom; (2) the atomic number limited to 100 (elements above fermium are assigned the one-hot-encoded position 101); (3) the formal charge; (4) the hybridization state; (5) the chiral tag representing, if applicable, the chiral type such as tetrahedral or octahedral; (6) whether the atom is part of a ring; and (7) the total number of hydrogens bonded to the atom. Figure 1b shows the bond invariants, namely: (1) the bond type, (2) the stereochemistry of the bond, (3) whether the bond is part of an aromatic system, (4) whether the bond is part of a conjugated system, (5) the type of the first atom connected by the bond, and (6) the atomic numbers of the two atoms forming the bond.

### Choice of set representation and GNN layers

Initially, we chose the set representation layer for MRS1 (which consists of only the set representation layer after a non-learned embedding based on atom invariants) by comparing the three architectures DeepSets, Set-Transformer and RepSet[14–16]. The hyperparameters were chosen based on defaults and findings from the initial publications of each method.

On the basis of the performance of the three different implementations of MSR1 (Extended Data Table 1) on the six datasets BACE, BBBP, ClinTox, ESOL, FreeSolv and Lipo, which were chosen for their relatively small size resulting in fast learning, we selected RepSet as our set representation layer of choice.

For the GNN layer, we implemented and evaluated GIN, GAT and GCN as graph embedding layers for our set representation extended graph neural network[7,18,45]. Again, we used default hyperparameters based on the original publications and, pairing each set representation layer with each GNN layer, evaluated different versions of our set representation-enhanced GNN. Using the same selection of benchmark datasets as with the set-representation layer selection, we evaluated the nine combinations of GNN and set representation layers (Extended Data Tables 2 and 3).

Overall, the results point towards a need for hyperparameter optimization depending on the combination of models. We therefore picked the combination not on the performance shown in Extended Data Tables 2 and 3 but on the fact that we use RepSet in MSR1 and MSR2, and that GIN(E) is generally reported as a well-performing baseline architecture in chemistry tasks[8,40,46].

### Hyperparameter optimization

For the graph-based models (the GINE baseline and SR-GINE), we ran an initial simple search using the BBBP dataset over both models while keeping the set representation parameters for SR-GINE fixed at eight hidden sets with eight elements each. For each set of hyperparameters, six models were trained on random seeds and their performance averaged. The best models were selected for lowest cross-entropy loss during validation. During this search, SR-GINE performed consistently better than the baseline (GINE). In a next step, we conducted a further simple grid search on the number of hidden sets and elements, as the

authors of the original publication have shown the performance to be influenced by these parameters[16]. We chose 128 hidden sets with 64 elements each based on tests on the BBBP dataset.

### Computational cost and scalability

We empirically investigated the SR-GINE approach's scalability by comparing its training time and performance with the non-extended GINE architecture. The average runtime for training SR-GINE on the OCELOT chromophore dataset ($N_{train}$ = 20,201) is 9 min 44 s ± 44 s, which is an increase of 6.2% compared with GINE (9 min 10 s ± 42 s). On the larger QM9 dataset ($N_{train}$ = 107,108), the training time increases to 50 min 25 s ± 0 min 47 s for GINE and 62 min 25 s ± 3 min 27 s for SR-GINE, an increase of 23.8%. In the case of QM9, this increase in training time is close to the average increase in performance, which is 24.4% (Extended Data Table 4). Although these numbers suggest that SR-GINE can scale well in terms of a performance–training time trade-off, it is not guaranteed that this is the case for all datasets. However, our experiments throughout this study showed a substantial increase of performance of SR-GINE over GINE on small- to medium-sized datasets where the increase in training time was marginal, as shown with the example of OCELOT.

Overall, these observations agree with the evaluation of RepSet by ref. 16. The timing data were taken from training runs on a Nvidia RTX 4090Ti graphics processing unit with 12 GB random-access memory for GINE and SR-GINE.

### Datasets and preprocessing

**MoleculeNet datasets.** The datasets that are often collected under the name MoleculeNet datasets, namely, HIV, BACE, BBBP, Tox21, SIDER, ClinTox, ESOL, FreeSolv, Lipo, QM7 and QM8, were split into train, validation and test sets based on Murcko scaffolds in accordance with refs. 17,20. Four other datasets from the MoleculeNet collection that are often used (ToxCast, MUV, PCBA, QM9) were omitted from the benchmarks, as they are either too large or have too many tasks to be processed with limited compute.

**OCELOT.** The OCELOT dataset was used, in accordance with the original publication[24], with a fivefold split and the subsequential five training–test iterations. In addition, a randomly sampled fraction of 0.2 of the training set was used as a validation set to facilitate early stopping.

**Biogen ADME assay data.** Reference [31] released their data with a predefined training/test split. No further processing of the data was necessary.

**PDBbind.** For PDBbind, the splits suggested by ref. 39 were used. The complexes found in the PDBbind core set were removed from the PDBbind refined set (of which core is a subset). PDBbind refined was subsequently used as a training set and PDBbind core as a test set.

**Buchwald–Hartwig (HTE).** The data and splits were provided by ref. 47 in their original publication. Comparison results were taken from ref. 43. No further processing of the data was necessary.

**Buchwald–Hartwig (ELN).** The data and splits were provided by ref. 48. No further processing of the data was necessary.

### Finetuning MolFormer

For our experiments that include MolFormer benchmark data that were not available in the original publication, we finetuned the model in the respective datasets. Specifically, we finetuned the available pretrained MolFormer model that has been made available publicly (checkpoint N-Step-Checkpoint_3_30000.ckpt) on the Biogen ADME and OCELOT datasets. The ADME and OCELOT models were finetuned for 500 and 300 epochs, respectively. For OCELOT, the number of epochs was reduced as we observed early plateauing. All other hyperparameters were chosen as suggested by ref. 21. Note that the finetuning required a substantial increase in hardware resources over the training of the graph- and set-based models due to a high memory requirement.

## Data availability

All data required to reproduce the results of this study, with the exception of the PDBbind dataset, have been made available in the GitHub repository https://github.com/daenuprobst/molsetrep (ref. 49). The PDBbind data can be downloaded from http://www.pdbbind.org.cn after a free registration process.

## Code availability

All code required to reproduce the results of this study as well as the source code of the pip package is available under the MIT License in the GitHub repository https://github.com/daenuprobst/molsetrep (ref. 49).

## References


1. Hamilton, W. L., Ying, R. & Leskovec, J. Representation learning on graphs: methods and applications. Preprint at http://arxiv.org/abs/1709.05584 (2018).
2. Filipa de Almeida, A., Moreira, R. & Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **3**, 589–604 (2019).
3. Walters, W. P. & Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.* **38**, 143–145 (2020).
4. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
5. Gupta, R. et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Divers.* **25**, 1315–1360 (2021).
6. Choudhary, K. et al. Recent advances and applications of deep learning methods in materials science. *NPJ Comput. Mater.* **8**, 59 (2022).
7. Hu, W. et al. *Proc. 8th International Conference on Learning Representations* (OpenReview.net, 2020).
8. Fang, X. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).
9. Zang, X., Zhao, X. & Tang, B. Hierarchical molecular graph self-supervised learning for property prediction. *Commun. Chem.* **6**, 34 (2023).
10. Axelrod, S. & Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci. Data* **9**, 185 (2022).
11. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
12. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
13. Jena, S. et al. Noncovalent interactions in proteins and nucleic acids: beyond hydrogen bonding and π-stacking. *Chem. Soc. Rev.* **51**, 4261–4286 (2022).
14. Zaheer, M. et al. Deep sets. In *Proc. 31st International Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 3394–3404 (Curran Associates Inc., 2017).
15. Lee, J. et al. Set transformer: a framework for attention-based permutation-invariant neural networks. *In Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. and Salakhutdinov, R.) 3744–3753 (PMLR, 2019).
16. Skianis, K., Nikolentzos, G., Limnios, S. & Vazirgiannis, M. Rep the set: neural networks for learning set representations. In *Proc. 23rd International Conference on Artificial Intelligence and Statistics* (eds Chiappa, S. and Calandra, R.) 1410–1420 (PMLR, 2020).

17. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
18. Veličković, P. et al. *Proc. 6th International Conference on Learning Representations* (OpenReview.net, 2018).
19. Gasteiger, J., Gros, J. & Günnemann, S. *International Conference on Learning Representations* (OpenReview.net, 2020).
20. Wang, Y., Wang, J., Cao, Z. & Farimani, A. B. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
21. Ross, J. et al. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.* **4**, 1256–1264 (2022).
22. Schütt, K. T., Gastegger, M., Tkatchenko, A., Müller, K.-R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
23. Dral, P. O. Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett.* **11**, 2336–2347 (2020).
24. Bhat, V. et al. Electronic, redox, and optical property prediction of organic π-conjugated molecules through a hierarchy of machine learning approaches. *Chem. Sci.* **14**, 203–213 (2022).
25. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 1263–1272 (PMLR, 2017).
26. Wenzel, F. et al. Assaying out-of-distribution generalization in transfer learning. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 7181–7198 (Curran Associates Inc., 2022).
27. Bao, Q. et al. A systematic evaluation of large language models on out-of-distribution logical reasoning tasks. Preprint at https://arxiv.org/abs/2310.09430v3 (2023).
28. Balani, S. K., Miwa, G. T., Gan, L.-S., Wu, J.-T. & Lee, F. W. Strategy of utilizing in vitro and in vivo ADME tools for lead optimization and drug candidate selection. *Curr. Top. Med. Chem.* **5**, 1033–1038 (2005).
29. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* **55**, 263–274 (2015).
30. Bhhatarai, B., Walters, W. P., Hop, C. E. C. A., Lanza, G. & Ekins, S. Opportunities and challenges using artificial intelligence in ADME/Tox. *Nat. Mater.* **18**, 418–422 (2019).
31. Fang, C. et al. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: an industrial perspective. *J. Chem. Inf. Model.* **63**, 3263–3274 (2023).
32. Mandal, S., Moudgil, M. & Mandal, S. K. Rational drug design. *Eur. J. Pharmacol.* **625**, 90–100 (2009).
33. Guvench, O. & MacKerell, A. D. Computational evaluation of protein–small molecule binding. *Curr. Opin. Struct. Biol.* **19**, 56–61 (2009).
34. Ballester, P. J. & Mitchell, J. B. O. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**, 1169–1175 (2010).
35. Crampon, K., Giorkallos, A., Deldossi, M., Baud, S. & Steffenel, L. A. Machine-learning methods for ligand–protein molecular docking. *Drug Discov. Today* **27**, 151–164 (2022).
36. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).
37. Hassan-Harrirou, H., Zhang, C. & Lemmin, T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *J. Chem. Inf. Model.* **60**, 2791–2802 (2020).
38. Meli, R., Anighoro, A., Bodkin, M. J., Morris, G. M. & Biggin, P. C. Learning protein–ligand binding affinity with atomic environment vectors. *J. Cheminform.* **13**, 59 (2021).
39. Li, S. et al. *Proc. 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, 2021).
40. Nguyen, T. et al. GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140–1147 (2021).
41. Zhang, S., Liu, Y. & Xie, L. A universal framework for accurate and efficient geometric deep learning of molecular systems. *Sci. Rep.* **13**, 19171 (2023).
42. Sánchez-Cruz, N., Medina-Franco, J. L., Mestres, J. & Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* **37**, 1376–1382 (2021).
43. Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit. Discov.* **1**, 91–97 (2022).
44. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform.* **12**, 43 (2020).
45. Kipf, T. N. & Welling, M. *Proc. 5th International Conference on Learning Representations* (OpenReview.net, 2017).
46. Peng, Y. et al. Enhanced graph isomorphism network for molecular ADMET properties prediction. *IEEE Access* **8**, 168344–168360 (2020).
47. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
48. Saebi, M. et al. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **14**, 4997–5005 (2023).
49. Probst, D. daenuprobst/molsetrep: release for publication. *Zenodo* https://doi.org/10.5281/zenodo.11148702 (2024).


## Acknowledgements

This work was supported by the Graph Neural Networks for Explainable Artificial Intelligence ERA-NET + EJP (20CH21_195579) and Integrated multiscale analysis of translation: single-molecules, omics and computation (CRSII5_205884) grants (P.V.). We thank F. Craighero, A. Hariri and N. Aspert for their valuable technical assistance and discussions on the implementation.

## Author contributions

D.P. and P.V. supervised the research. M.B. and D.P. conceived the research. M.B. and D.P. performed the experiments and wrote the Python scripts. M.B. analysed the experimental data. M.B. and D.P. wrote the paper.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s42256-024-00856-0.

**Correspondence and requests for materials** should be addressed to Daniel Probst.

**Peer review information** *Nature Machine Intelligence* thanks Rocío Mercado and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s42256-024-00856-0

**Extended Data Table 1 | Comparison between Deep Sets, Set-Transformer, and RepSet on a subset of the MoleculeNet benchmark data sets**

| Data Set | Metric | DeepSets | Set-Transformer | RepSet |
|---|---|---|---|---|
| BACE | AUROC | 76.4±2.3 | 55.4±1.9 | **77.2±3.6** |
| BBBP | AUROC | 71.1±1.0 | 69.9±3.3 | **73.4±0.7** |
| ClinTox | AUROC | 82.8±1.6 | 83.3±4.6 | **85.6±1.3** |
| ESOL | RMSE | 0.632±0.086 | 0.731±0.046 | **0.584±0.019** |
| FreeSolv | RMSE | 2.542±0.060 | **2.144±0.122** | 2.154±0.515 |
| Lipo | RMSE | 1.228±0.076 | 1.080±0.038 | **0.884±0.028** |

Training and testing were run 3 times; mean and standard deviation are reported. Values highlighted in bold represent the best performance per data set.

https://doi.org/10.1038/s42256-024-00856-0

**Extended Data Table 2 | Classification benchmarks on different GNN-set representation combinations, where the set representation layer replaces global pooling**

| | BACE | | | BBBP | | | ClinTox | | |
|---|---|---|---|---|---|---|---|---|---|
| | GINE | GAT | GCN | GINE | GAT | GCN | GINE | GAT | GCN |
| RepSet | 63.9±25.1 | **78.6±0.5** | 76.2±2.0 | 67.2±3.5 | 67.8±2.8 | **68.1±3.1** | 83.3±3.4 | 83.2±3.9 | **87.0±2.7** |
| Set-Transformer | **66.0±2.1** | 65.2±2.3 | 63.8±1.0 | 70.2±1.2 | 67.9±2.2 | **73.0±1.3** | **84.1±5.5** | 76.9±6.1 | 83.8±6.8 |
| Deepsets | 62.1±3.2 | 64.1±0.9 | **64.7±1.4** | **71.0±1.0** | **71.0±3.8** | 69.4±1.4 | 87.3±3.0 | 86.5±3.2 | **88.3±2.8** |

Training and testing were run 3 times; mean and standard deviation are reported. The metric used is AUROC. The values in bold represent the best performance by set representation method and data set, the underlined values represent the best performance per GNN method and data set.

https://doi.org/10.1038/s42256-024-00856-0

**Extended Data Table 3 | Regression benchmarks on different GNN-set representation combinations, where the set representation layer replaces global pooling**

| | FreeSolv | | | ESOL | | | Lipo | | |
|---|---|---|---|---|---|---|---|---|---|
| | GINE | GAT | GCN | GINE | GAT | GCN | GINE | GAT | GCN |
| RepSet | 3.268±0.081 | 2.615±0.242 | **2.356±0.067** | 0.703±0.034 | **0.522±0.028** | 0.637±0.039 | 0.841±0.010 | **0.786±0.027** | 0.800±0.022 |
| Set-Transformer | 3.153±0.185 | 2.216±0.058 | **2.084±0.135** | 0.746±0.028 | **0.626±0.052** | 0.685±0.054 | 0.871±0.030 | **0.753±0.057** | 0.791±0.043 |
| Deepsets | 3.127±0.182 | 2.401±0.172 | **2.274±0.257** | 0.729±0.088 | **0.693±0.076** | 0.699±0.089 | 0.795±0.015 | 0.800±0.032 | **0.784±0.047** |

Training and testing were run 3 times; mean and standard deviation are reported. The metric used is RMSE. The values in bold represent the best performance by set representation method and data set, the underlined values represent the best performance per GNN method and data set.

https://doi.org/10.1038/s42256-024-00856-0

**Extended Data Table 4 | Empirically evaluating the scalability of extending GINE with a set representation layer on the QM9 data set ($N_{train}$ = 107108)**

| Property | $\varepsilon_{HOMO}$ | $\varepsilon_{LUMO}$ | $\Delta\varepsilon$ | ZPVE | $\mu$ | $\alpha$ | $\langle R^2 \rangle$ | $C_v$ |
|---|---|---|---|---|---|---|---|---|
| Unit | eV | eV | eV | eV | D | $\text{bohr}^3$ | $\text{bohr}^2$ | cal/mol K |
| D-MPNN | **0.093±0.005** | 0.106±0.002 | 0.148±0.003 | 0.037±0.004 | **0.450±0.006** | 0.493±0.008 | **24.371±0.922** | 0.244±0.005 |
| GINE | 0.106±0.002 | 0.104±0.002 | 0.151±0.004 | 0.054±0.005 | 0.515±0.006 | 0.685±0.007 | 31.870±1.208 | 0.428±0.004 |
| SR-GINE | 0.101±0.001 | **0.100±0.001** | **0.140±0.004** | **0.015±0.000** | 0.491±0.002 | **0.487±0.922** | 25.852±0.448 | **0.194±0.006** |

While the training time is increased by 23.8%, the predictive performance is increased by 24.4%. In addition, SR-GINE performs better overall than the D-MPNN baseline. The metric used is MAE. The value representing the best performance for each task is highlighted in bold.