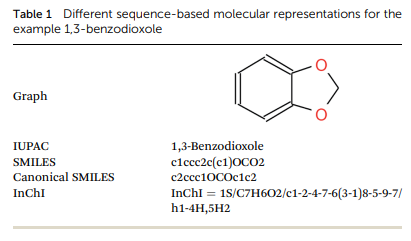
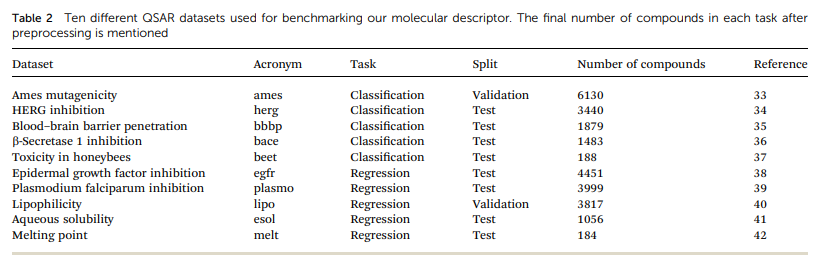
# Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations

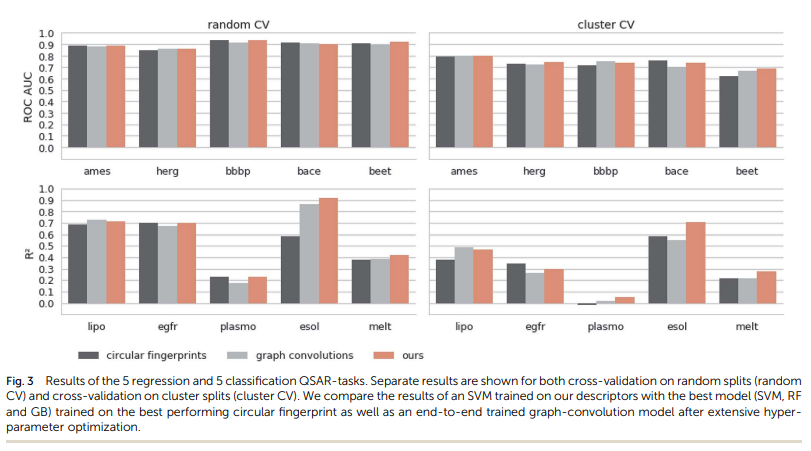
* The researchers purpose a method that exploit the powerful ability of deep neural networks to learn a feature representation from low-level encodings of a huge corpus of chemical structures.
* It translates between two semantically equivalent but syntactically different representation of molecular structures, compressing the meaningful information both representations have in common in a low-dimensional representation vector.
* Introduction: - Molecular fingerprints encode structural or functional features of molecules in a bit string format and are commonly used for tasks like virtual screening, similarity searching and clustering.
* The circular fingerprint like the *extended-connectivity fingerprints* (ECFPs) were introduced to model quantitative structure-activity relationships (QSAR) for biological endpoints by way of classical machine learning approaches as well as for ligand-based virtual screening (VS).
* The purposed DNNs have in common that they use pre-extracted molecular descriptors (mostly ECFPs) as input features.
* By training a DNN directly on a comprehensive and low-level representation, it can automatically learn to extract useful descriptors best suited for the specific task it is trained on, resulting in a specific descriptor set for a given dataset.
* Because it learns from scratch for every new dataset, there is a high chance of overfitting.
* Gomez-Bombarelli proposed a *variational autoencoder* to convert the discrete SMILES representation.
* They show that the representations could also be used as descriptors for a downstream classification task. Xu proposed a related unsupervised approach based on *sequence to sequence learning.*
* In this the encoder has input as SMILES sequence of variable length with discrete values.
* The autoencoder network is trained on minimizing the mean reconstruction error on a single-character level for each input sequence.
* By introducing an *information bottleneck* between the encoder and the decoder, the network if forced to compress the essential information of the input, so the decoder still makes as few errors as possible in the reconstructions.
* There is risk with autoencoder that an autoencoder on reconstructing a sequence which represents a molecule bears the risk that the network solely focuses on syntactic features and repetitive patterns of this sequence, neglecting its semantics and failing to encode higher-level concepts such as molecular properties.
* Neural Machine Translation (NMT) model first reads the whole input sequence and encodes it into an intermediate continuous vector representation (latent representation) which is then used by the decoder to emit a respective translation.
* The researchers want to exploit translation methodology to extract the “meaning” of a molecular representation like an InChI by translating it to another syntactically different one e.g., SMILES.
* The researchers propose a model that can extract the information contained in a comprehensive but discrete and variable-sized molecular representation and transform it into a continuous and fixed-sized representation.
* Methods: - molecular representation: - the InChI notation represents molecular structures as a sequence of characters divided into layers and sub-layers providing different types of information such as the chemical formula, bonds and charges.



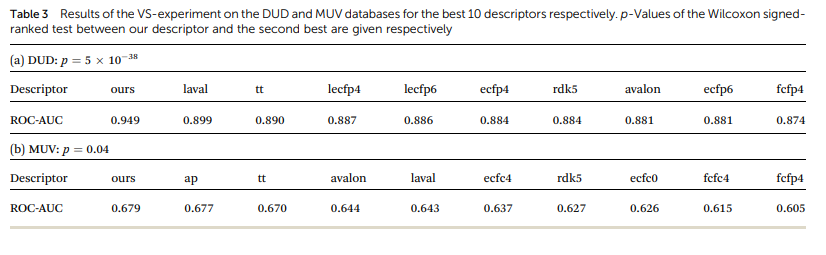
* In the above diagram we can see that SMILES and Canonical SMILES share the same identifiers and general syntax, but they are not identical.
* The researchers tokenized the sequences and encoded them in a one-hot vector representation.
* A lookup table T for the *N* tokens in sequence representations (e.g., T2 = C, T5 = Br), a one-hot representation of token Ti is defined by an *N*-dimensional vector with a one in the *i*-th entry and zeros elsewhere.
* Translation model: - for the encoder network, the researcher tried both convolutional neural network (CNN) and recurrent neural network (RNN) architectures of different size and depth followed by a fully connected layer that maps the output of the CNN or the concatenated cell states of the RNN to the latent space, respectively.
* The researchers also extend the translation model by an additional classification model for certain molecular properties.
* Like Gomez-Bombarelli method, this classification model takes the latent representation of the translation model as the input and predicts certain molecular properties which can be directly deduced from the molecular structure.
* The complete model is trained on minimizing the cross-entropy between these probability distributions and the one-hot encoded correct characters in the target sequence as well as minimizing the mean squared error in predicting the molecular properties.
* For decoder RNN they utilized *teacher forcing* during training and a left-to-right beam search during inference.
* To select the best combination of translation task and architecture, they used the predictive performance of machine learning models built on two QSAR datasets using the respective latent representations as descriptors.
* Datasets and preprocessing: - the translation model was pretrained from ZINC15 and PubChem datasets. They take only those data who falls under these criteria: - only organic molecular weight between 12 and 600, more than 3 heavy atoms and a partition coefficient log *P* between – 7 and 5.
* For each molecule, nine molecular properties were extracted: log *P*, the maximal and minimal partial charge, the number of valence electrons, the number of hydrogen bond donors and acceptors, Balaban’s *J* value, the molar refractivity and the topological polar surface area.
* For the final translation model, they performed eight QSAR and two VS experiments. Two of the datasets were used to validate the different translation models’ architectures. The remaining eight datasets were solely used for evaluating the final model.



* Evaluation and baseline: - the researchers compare three different approaches: - classical machine learning models applied on descriptors and on circular fingerprints of different radii and folding as implemented in RDKit as well as an end-to-end molecular graph convolution method as implemented in DeepChem.
* The first two model uses Random Forest (RF), support vector machine (SVM) with an RBF kernel and Gradient Boosting (GB) as implemented in scikit-learn.
* The SVM method work best with the combination researchers’ descriptors and therefore applied to all other QSAR datasets for their descriptors.
* The researchers perform an extensive hyperparameter optimization in a nested cross-validation (CV) fashion to select the best set of descriptors, model and hyperparameters for each task.
* In this each dataset was split into two different ways for the validation. The random CV corresponds to five random splits while the cluster CV corresponds to five clusters obtained by *K­*-means clustering with *K*=5 on MACCS fingerprints.
* For selecting the best combination, they specifically looked at the coefficient of determination (*r2*) and the area under the receiver operating characteristic curve (ROC AUC) for the regression and classification tasks.
* For the ligand-based virtual screening experiments, they followed the benchmarking protocol proposed by Riniker for each target in both VS databases, five active compounds were picked randomly and the remaining compounds were ranked according to their similarity to the active set as measured by a similarity metric in the respective descriptor space.
* The similarity in the discrete baseline fingerprint space was calculated using the Tanimoto similarity. For continuous descriptors cosine was used and for calculating the resulting ranking of the compounds the mean ROC-AUC over the 50 repetitions for each target was used.
* Results and Discussion: - the translation model is data-driven model that generates meaningful compound representations by forcing translation of all necessary information between two sequence-based representations of a molecule into a low dimensional continuous embedding.
* After pretraining is finalized, the translation model can be used to encode compounds into the embedding or to decode embeddings into compounds.
* Pretraining: - the best performing model for both translation tasks as well as the best model for the regular canonical SMILES autoencoding task.
* The translation model is forced to store all important information necessary to do the translation in the bottleneck of the network: the latent representation.
* The overall best performance was achieved with a translation model based on an RNN architecture for the encoder network that was trained on translating from a SMILES representation to its canonical version.
* The additional classification task seems to have a clear positive impact on the predictive performance of the lipophilicity task, while resulting in a small improvement on the Ames mutagenicity task.
* Translating between two molecular representations seems harder to learn than reconstructing the same input representation.
* The translation models cannot simply store sequence-based features or patterns in the latent space, but have to learn to extract the information that both the input and output sequences have in common: the molecule they are both representing.
* QSAR modelling: - the below figure shows the results of evaluation for random-split and cluster-split cross-validation respectively, comparing the researcher molecular descriptor to the best model based on the different circular fingerprints and the graph-convolution networks trained end-to-end for each QSAR dataset individually.



* The graph-convolution method outperforms the models based on the baseline fingerprint in predicting these physico-chemical endpoints. In the case of cluster splits, the graph-convolution model apparently fails to generalize on the hold-out clusters.
* The graph-convolution method is trained end-to-end, it has to learn to extract meaningful features for each dataset from scratch which could lead to overfitting, if training data are limited.
* The researchers fixed their feature extraction method based on two datasets (Ames and lipophilicity on random splits) to avoid a model selection bias on the remaining test sets.
* The fingerprint-based models could choose between nine different flavours of circular fingerprints and three different learning algorithms for each task respectively and due to the considerable training time, the graph-convolution models were not trained in a nested cross-validation.
* Virtual screening: - the goal of ligand-based virtual screening (VS) is to rank a large set of compounds with respect to their activity on a certain target based on their similarity to some known active query compounds.
* They follow the benchmark protocol of Riniker to compare their descriptors against other state-of-the-art molecular descriptors.
* They use DUD and MUV dataset and the descriptor outperformed the second-best descriptor (*p*<0.05).



* Exploring the continuous descriptor space: - the proposed descriptor is continuous and the encoding into the descriptor space is reversible, due to the decoder part of the translation model.
* Gomez-Bombarelli already shows that a continuous encoding of a molecular structure enables us to explore the neighborhood of this molecule by decoding from points close to the query molecules’ embedding.
* The incremental shifts in the continuous descriptor space correspond to smooth transitions in the discrete chemical space.
* The first principal component of their pretraining dataset correlates with the size of molecules: adding or subtracting a value along this axis corresponds to adding or removing atoms from the structures.
* The second principal component of the pretraining dataset seem to be correlated with altering the molecule’s polarity.
* The mean Spearman correlation coefficient *r* between the compounds’ molar weight and the respective step along the first principle component was *r* = 0.9470 (*p* = 0.00048).
* The mean correlation between the compound’s partition coefficient log *P* and the respective step along the second principle component was *r* = -0.916 (*p* = 0.00015).
* The researcher observed that there is a clear correlation between the (Euclidean) distance in their descriptor space and the (Tanimoto) distance in the circular fingerprint space.
* They find that if the most probable output of the model’s beam search decoder results in an invalid SMILES, they observe that it is likely that one of the next most probable sequences results in a valid SMILES (>99%).
* In a similar study Blaschke analyzed 4 different autoencoder frameworks on the SMILES-to-SMILES reconstruction task and reported a valid SMILES proportion of only approximately 20% using their best model, if moved away by a similar distane.
* Conclusion: - the researcher proposed a novel methodology that is able to learn to extract meaningful molecular descriptors, solely by an unsupervised training on a large dataset of molecular structures.
* They show that molecular descriptors extracted by their method significantly outperform state-of-art molecular fingerprints in a ligand-based virtual screening (VS) experiment. They also show that ML models based on their descriptor perform similar to various quantitative structure-activity relationships (QSAR tasks), compare to multiple state-of-the-art molecular fingerprints and computationally expensive graph-convolution models.
* The researcher’s model learns its own extraction method in the data-driven way.
* The follow-up this model will be the translation of conceptually different molecular representations such as the molecular graph or 3D-structure-based representations like the van der Waals and/or electro-negative potential surface.
* They observe smooth and meaningful translations in the chemical structure when a molecule’s embedding is shifted in certain directions, where shifts along different axes in their descriptor space correspond to different structural and functional properties in the chemical space.