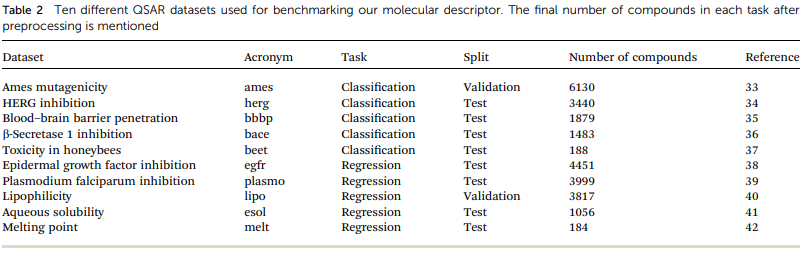
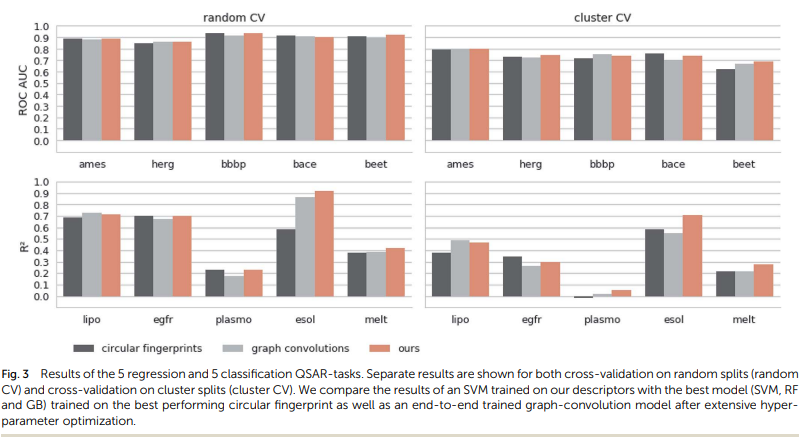
# Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations

* Abstract: - the recent surge of interest in using ML across chemical space make the design and predict molecular properties easy. This work relies on defining clever feature representations, in which the chemical graph structure is encoded in a uniform way such that predictions across chemical space can be made.
* In this paper the researchers propose to exploit the powerful ability of deep neural networks to learn a feature representation from low-level encodings of a huge corpus of chemical structures.
* Their model borrows ideas from neural machine translation: it translates between two semantically equivalent but syntactically different representations of molecular structures, compressing the meaningful information both representations have in common in a low-dimensional representation vector.
* The researchers’ model shows competitive performance in modelling quantitative structure-activity relationships in all analyzed datasets.
* The continuity of the descriptor space and the existence of the decoder that permits deducing a chemical structure from an embedding vector allow for exploration of the space and open up new opportunities for compound optimization and idea generation.
* Introduction: - a widely used concept of generate such theoretical molecular descriptors is *molecular fingerprints.*
* Molecular fingerprints encode format and are commonly used for tasks like virtual screening, similarity searching and clustering.
* *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 advantage of DNNs proposed that in common they can use pre-extracted molecular descriptors as input features.
* It contradicts the fundamental idea of representation learning: DNNs should learn a suitable representation of the data from a simple but complete featurization, rather than relying on sophisticated human-engineered representations.
* 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 features have to be learned from scratch for every new dataset, these methods are prone to overfitting if trained on limited data. It is a major issue when it comes to bioactivity data, due to the relatively high cost of generating a data point.
* Gómez-Bombarelli et al. proposed a *variational autoencoder* to convert the discrete SMILES representation of a molecule to and from a multidimensional continuous representation.
* Xu et al. proposed a related unsupervised approach based on *sequence-to-sequence learning.*
* Both studies use *autoencoder*, the whole 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 is forced to compress the essential information of the input, so that the decoder still makes as few errors as possible in the reconstruction.
* Training 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.
* In this work, the researchers want to address this issue by proposing a method that is based on a translation rather than a reconstruction methodology.
* A so-called Neural Machine Translation model first reads the whole input sequence and encodes it into an intermediate continuous vector representation which is then used by the decoder to emit a respective translation.
* In this the researchers wants to exploit the translation methodology to extract the “meaning” of a molecular representation like an InChI (International Chemical Identifier) by translating it to another syntactically different one, *e.g.,* SMILES.
* The decoder can only succeed in generating the high translation for a given molecular representation if the encoder compresses a comprehensive description of the chemical structure in the latent representation.
* Once trained, the resulting model can be used to extract meaningful molecular descriptors for query structures without the need for retraining or including labels.
* At last, the researchers show that it is possible to navigate smoothly in this new continuous chemical descriptor space by modifying slightly the molecular representation of an existing compound in a given direction and using the decoder to obtain new chemical structures.
* Methods: - molecular representation – in this the researchers focuses on the sequence-based SMILES and InChI representations.
* 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.
* The SMILES notation also represents molecular structures as a sequence of characters. But a SMILES is not divided into different information layers but encodes the whole molecular structure in one sequence of characters including identifiers for atoms as well as identifiers denoting topological features like bonds, rings and branches.
* The different notations differ in their syntax while representing the same molecule, like SMILES and canonical SMILES share the same identifiers and general syntax the two sequences, coming from different algorithm, are not identical.
* The researchers utilized the SMILES enumeration procedure proposed by E. Bjerrum to generate a random SMILES variant for a given molecule. In order to be invariant to the SMILES representation at inference time, they also used the canonical SMILES as the input half of the time.
* To use the aforementioned sequence-based molecular representations as the input and output of the translation model, the researchers tokenized the sequences and encoded them in a one-hot vector representation.
* By defining a lookup table T for the *N* tokens in sequence representations (*e.g.,* T2 = C, T5 =Br), a one-hot representation of token is defined by *N*-dimensional vector with a one in the *i*-th entry and zeros elsewhere.
* They defined different lookup tables for both SMILES and InChI representations, they tokenized 38 and 28 unique characters for SMILES and InChI sequences, respectively.
* Translation model – the researchers 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.
* In this approach the decoder network consists of an RNN, whose cell states are initialized by an individual fully connected layer for each layer in the RNN, taking the latent space as the input.
* To enhance the models’ performance, they extend the translation model by an additional classification model for certain molecular properties.
* 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 structures.
* The output of the decoder networks’ RNN is a sequence of probability distributions over the different possible characters defined.
* The whole 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.
* The decoder 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 on a large dataset composed of molecular structures from the ZINC15 and PubChem databases.
* Both datasets merged and filter with RDKit using these criteria – only organic molecules, 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 electron, 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. The QSAR datasets were taken from various sources and were preprocessed in the same way as the pretraining dataset. Two of the datasets were used to validate the different translation models’ architectures. The remaining eight datasets were solely used for evaluation the final model.
* The VS experiment were performed on 40 targets of the Directory of Useful Decoys (DUD) and 17 targets of the Maximum Unbiased Validation (MUV) dataset.



* Evaluation and baseline – for modeling structure-activity relationships, the researchers compare three different approaches: classical machine learning models applied on their 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.
* In this method the first two methods require selecting the learning algorithm to plug on top of the molecular representation.
* For the first two methods the researchers used Random Forest (RF), support vector machine (SVM) with an RBF kernel and Gradient Boosting (GB) as implemented in scikit-learn.
* The SVM works best in all the algorithms and therefore the only method applied to all other QSAR datasets for researchers’ descriptors.
* They performed an extensive hyperparameter optimization in a nested cross-validation (CV) fashion to select the best set of descriptors, model and hyperparameters for each task.
* The graph convolution models were trained directly on the different QSAR datasets, learning rate and filter size were optimized in a cross-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.
* To select the best performing combinations, 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, respectively.
* For the ligand-based virtual screening experiments, they followed the benchmark protocol proposed by Riniker et al for each target in both VS databases, five active compounds were picked randomly and the remaining 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.
* This process continuous 50 times and every time it selects a new random set of active and decoy compounds.
* The similarity of the discrete baseline fingerprint space was calculated using the Tanimoto similarity. For their continuous descriptors they used cosine similarity.
* The resulting ranking of the compounds is evaluated by calculating the mean ROC-AUC over the 50 repetitions for each target.
* A Wilcoxon signed-rank test is performed to analyzed the statistical significance of the differences in the mean ranks of our descriptor to the baseline descriptors.
* Results and discussion: - the model is a data-driven method, it translates all necessary information between two sequence-based representations of a molecule into a low dimensional continuous embedding.
* Since sequence-based representations of molecules such as SMILES or InChI are easily obtained from cheminformatics packages, the pretraining of the model can be performed on a vast chemical space.
* After the pretraining is finalized, the translation model can be used to encode compounds into the embedding or to decode embeddings into compounds.
* Pretraining – the researchers evaluate the different network architectures of the translation model in terms of performance of the extracted descriptors for the two validation tasks.
* They show the best performing model for both translation tasks as well as the best model for the regular canonical autoencoding task.
* Commonly, as the models get better at translating the input to the output sequence, the predictive performance of an SVM based on the latent representation also improves.
* Translation model is forced to store all important information necessary to do this translation in the bottleneck of the network: the latent representation.
* The more the data translation model get the better it suited as a molecular descriptor to predict certain properties of the molecule.
* 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 researchers try to train models on translating from canonical SMILES to InChI representations. However, these models failed to learn anything.
* It may because the InChI is quite complex format and making the generation of a correct InChI string for a given molecule a difficult task to learn.
* Because this model only focuses on translating, it reaches better translation accuracies faster.
* 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.
* All models based on translating between two different molecular representations show a clear improvement over models trained on reconstructing the same input sequence.
* Translating between two molecular representations seems harder to learn than reconstructing the same input representation.
* The translation model 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 – after that the researchers extracted molecular descriptors of the remaining (test) QSAR datasets with the best performing translation model and benchmarked.
* The below figure shows the results of this evaluation for random-split and cluster-split cross-validation respectively, comparing their 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 hyperparameter-optimized methods perform on a comparable level for most of the QSAR tasks, each method showing at least on one task a slightly better mean performance over the different splits.
* The lipophilicity and aqueous solubility datasets show the largest variance in performance between the models.
* In the case of random split, the graph-convolution method outperforms the models based on the baseline fingerprint in predicting these physico-chemical endpoint.
* In the 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-convolution method is trained end-to-end, it has to learn to extract meaningful feature extraction methods respectively, independently from the task at hand.
* The researchers’ proposed molecular descriptors exhibit competitive or better performance than the best baseline models in all investigate QSAR tasks.
* The researchers fixed their feature extraction method based on two datasets 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 aim of the 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.
* The researchers follow the Riniker et al. to check their models benchmark, to compare their extracted descriptors against other state-of-the-art molecules descriptors.
* On both databases (DUD and MUV) their descriptor significantly outperformed the second-best descriptor (*p* < 0.05).
* The best baseline descriptor in the DUD screen (laval) is only fifth in the MUV screen. The best baseline descriptor in the MUV screen (ap) is not even represented in the top ten performing, descriptors in the DUD screen.
* Exploring the continuous descriptor space: - the researchers’ proposed descriptor is continuous and the encoding into the descriptor space is reversible, due to the decoder part of their translation model.
* Gómez-Bombarelli et al. already shown that a continuous encoding of a molecular structure enables researchers to explore the neighborhood of this molecule by decoding from points close to the query molecule’s embedding.
* They incrementally shift the embedding of a query molecule in two different directions and decode it back to a molecule.
* The directions they are shifting the molecule’s embedding along are defined by the first and second principal component of the pretraining dataset in their descriptor space.
* They observe that 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 structure.
* The second principal component of the pretraining dataset seem to be correlated with altering the molecules’ polarity.
* They repeated the experiment with 1000 randomly picked compounds from validation dataset and shift every one of them 10 steps in negative and positive direction along with two principle components.
* The mean Spearman correlation coefficient *r* between the compound’s 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).
* All analyzed points along these two axes, when decoded resulted in a valid SMILES.
* Further increase the random direction to 100 and 1000 randomly picked compounds shows that a clear correlation between the (Euclidean) distance in researchers descriptor space and the (Tanimoto) distance in the circular fingerprint space.
* Even over a long-distances, their model succeeds in generating a high proportion of valid SMILES (>97%).
* They also find that if 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%).
* Blaschke et al. also did similar study with 4 different autoencoder frameworks on the SMILES-to-SMILES reconstruction task they find only approximate 20% valid SMILES proportion right.
* 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 their molecular descriptors extracted by their model outperform state-of-the-art molecular fingerprints in ligand-based virtual screening (VS) experiments.
* They also show that machine learning models based on their descriptor perform similarly- if not better – on various quantitative structure-activity relationships (QSAR tasks), when compared to multiple state-of-the-art molecular fingerprints and computationally expensive graph-convolution models.
* Their model has the advantage of both baseline models, the model does not depend fixed feature extraction rules but learns its own extraction method in a data-driven way.
* The researchers observed smooth and meaningful transition in the chemical structure when a molecules’ embedding is shifted in certain directions, where shifts along different aces in their descriptor space correspond to different structural and functional properties in the chemical space.
* The researchers’ model’s latent space was shown to be significantly better correlated with the molecule’s biochemical properties, they think that their proposed translation method could significantly improve such a method’s ability to generate and optimize molecules, also enabling optimization with respect to biological activity.

Robin Winter, Floriane Montanari, Frank Noé and Djork-Arné Clevert