Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Sergey Sosnin
PhD Program: Computational and Data Science and Engineering
Title of Thesis: Exploration of chemical space by machine learning
Supervisor: Professor Maxim Fedorov

The thesis document includes the following changes in answer to the external review process.

Following Natalia Strushkevich's recommendations, I slightly revised the structure of the Thesis. I added to the corresponding chapters new sections: 4.1 and 5.1, ("Materials and Methods") I renamed Chapter 2 to "The Literature Review." I renamed "Additional Information" to "Supplementary Material." I made minor corrections in all chapters to fix typos and clarify some points that were hard to follow for the reviewers. I fixed Table 5.1 (one column was lost). Other minor corrections are presented below in response to the reviewer's remarks.

Major comments:

Chapter 3 (3D CNN and BCF):

(Nikolay Brilliantov) Why the author uses only data obtained by the 3D RISM simulations and does not apply available experimental data for the structure of the solvent shell around a solute molecule. This could be e.g. data from the mean residence time of water molecules in the shell, measured by NMR, data on thermal neutron scattering, etc. I can admit that the author strives to develop a regular approach, while such experimental data are very sparse. Nevertheless, the additional crosscheck with the experimental data would be very beneficial. At least I expect a short discussion in the text about the possibility of the usage of experimental data

The reviewer's assumption is right. It is hard to find standardized experimental data for small molecules. We were going to create a universal pipeline for which only chemical structures are required. Moreover, we demonstrate that our method's performance depends on the number of conformers for each compound we use. So we use multiple conformers, and I believe that none of them represents real conformations in solutions. I added a paragraph with a discussion about this problem to the end of section 3.2

(Nikolay Brilliantov) It is not clear, why the author believes that the structure of the solvation shell of a chemical substance is a reliable toxicity indicator. Is this a just hypothesis,

confirmed by the subsequent analysis? Or do there exist some regular studies which prove the importance of the solvation shell structure for the toxicity? In the latter case, the according references should be provided, and their content briefly discussed.

Bioconcentration is not toxicity. As described in section 3, it is a ratio of stationary concentrations of a compound inside and outside of an animal (generally, fish). But the compound itself doesn't need to be toxic. However, if a compound has high BCF, it can be concerning for environmental chemists. Typically, hydrophobic compounds penetrate aqueous organisms well and concentrate inside. Hydrophilicity and hydrophobicity of a compound correlate with a solvation shell, so it can explain why our methods work well on this particular endpoint. It was discussed in section 3.1, subsection "Bioconcentration factor."

(Nikolay Brilliantov) Why the graph convolution model shows notably worse results than the baseline model (US EPA)? Although it is difficult to formulate the reason rigorously, a discussion of possible explanations is very welcome.

This question requires a separate investigation however I can cite a review (Sun et al. Graph convolutional networks for computational drug development and discovery, Briefings in Bioinformatics, Volume 21, Issue 3, May 2020, Pages 919–935) "Deep architectures require a large amount of training data in order to achieve significant improvements in predictive power and it is common that some tasks may contain insufficient data to make meaningful predictions." In our experience, GCN networks demonstrate good performance on large datasets but EPA's BCF dataset is quite small. Moreover, GCN works only with structural information, ignoring 3D representations.

(Nikolay Brilliantov) The statement in page 51 "only conformers with mutual RMSD (computed on the heavy atoms) more than 0.5A have been kept" is very vague and is not well explained: What are "heavy atoms"? How do you define "mutual RMSD"? Please reformulate this piece of the text accordingly
(Peter Ertl) Quite sophisticated method was used to generate conformations, would not much simple approach using standard RDKit conformation generator provide results of the same quality?

In fact, we use RDKit for conformers generation. We regard conformers generation as a part of the data augmentation process. For that point, we need conformers that, on the one hand, low energy (to assure physical meaning for such conformations); on the other hand, these conformers should be diverse to provide more information for our 3D CNN. To generate conformers by RDKit, we make an initial guess (using rdkit.Chem.rdDistGeom.EmbedMultipleConfs) first, and then perform a forcefield optimization; for example, we use UFF forcefield. But after the forcefield optimization, some

molecules fall into very similar conformations. To restrict the number of "valuable" conformers, we do Butina clustering. Then, on each iteration, we select one compound from each cluster until we reach the number of conformers we need. But when we select a candidate, we calculate a set of RMSD overall non-hydrogen atoms (heavy atoms) between the candidate and all conformers we have already chosen. If there is a low RMSD value, we exclude this candidate and continue iteration. This, maybe quite a complicated procedure, prevents the final conformers from being similar to each other. Although we use RDKit machinery for the conformers generation, the process of conformer selection is not a part of the standard RDKit pipeline. Our experiments, and experiments of other researchers (10.1186/s13321-020-00420-z), demonstrated that the number of conformers directly affects the performance of modeling; however, we believe that similar conformers bring no useful information for the neural network. That is the reason why we use this quite sophisticated method.

(Daniel Svozil): I don't fully understand how "the merging of BCF values of different experiments is possible". Doesn't the dissimilarity between species imply differences in the distribution of the organic compound within them? Please, could you elaborate more on why is it that BCF of different species can be compared?

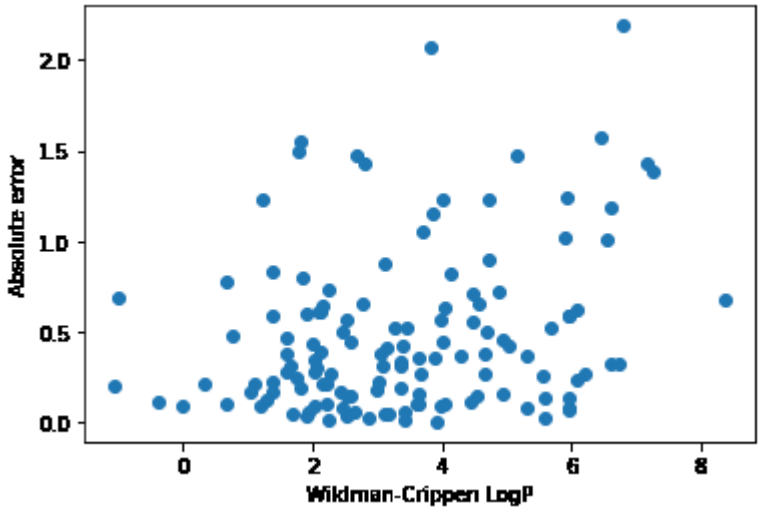
Yes, the metabolism and bioconcentration ability of chemical compounds differ in different organisms. If we discuss only fish species, because the US EPA dataset consists only of fish BCF values, it is known that it mostly depends on the fish species' lipid content (10.1080/10659360500474623, OECD 305 Guideline). OECD 305 Guideline indicates that "bioconcentration should be expressed as normalized to a fish with a 5% lipid content (based on whole body wet weight) in addition to that derived directly from the study. This is necessary to provide a basis from which results for different substances and/or test species can be compared against one another." One can speculate that this sentence presumes the existence of such a basis on which the results can be compared and probably merged.

Regarding the US EPA dataset, the agency does not describe how they collected the BCF database in detail, providing only links to other sources, so we can say nothing about whether this database was collected based on one species or many. Moreover, some of their sources are unavailable now (see ref. 74 in User's Guide for T.E.S.T. (version 4.2)). But if we have a look at the ref. 72, we see that this dataset consists of at least two species. So, in practice, it is admissible merging values for fish species, and, at least, US EPA has nothing contrary to it.

To avoid misinterpretation, I changed this sentence to: "However, OECD 305 guideline (Bioaccumulation in Fish: Aqueous and Dietary Exposure) allows comparing measurements even for different fish species under certain circumstances. "

(Daniel Svozil): Are there any strongly hydrophobic molecules in the test data set and did you check the predictions of your models for strongly hydrophobic molecules?

To explore this question, I built a correlation plot between LogP (calculated by rdkit) and absolute error for each compound in the test set. One can see that there is no strict correlation between these factors, however the error for compounds with high logP values are large. I added this plot and corresponding text to Section 3.1



Chapter 4 (Multitask learning for acute toxicity modelling)

(Peter Ertl) Chapter 4 "processing of intervals" this is well discussed topic in QSAR - such values are called normally qualified data or censored data, see for example
https://www.tandfonline.com/doi/abs/10.1080/15459621003609713
https://academic.oup.com/biomet/article-abstract/66/3/429/232342

I am thankful to the reviewer for this indication. Indeed, I did not know that this topic is well discussed in survival studies. However, I did not find any information about the DNN based modeling of censored data. As far as I can understand, in such studies, particular statistical models are used. Our approach was just a modification of the loss function, and as we mentioned, it does not succeed well. I believe that a separate study is required regarding the modeling of censored toxicity data by DNN. I added remarks to the subchapter “processing of intervals” and renamed it to “Censored data modeling.”

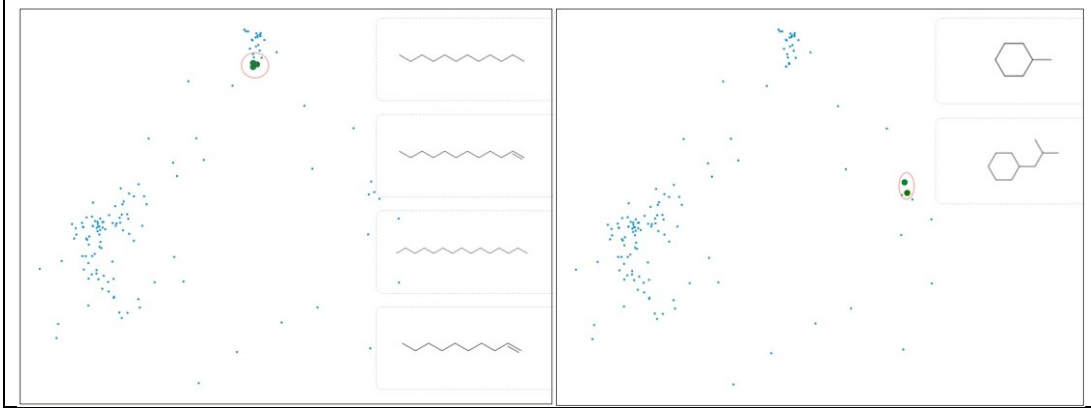
(Daniel Svozil) In this chapter, several claims about the superiority of MT_DNN models are made. In Conclusions part (page 75) you even claim that MT_DNN significantly (it doesn’t mean statistical significance, does it?) improve toxicity prediction. However, in my opinion, such bold claims should be supported by statistical model comparison. For example, you state that ST_DNN models are comparable with XGBoost and RF but, based on the visual

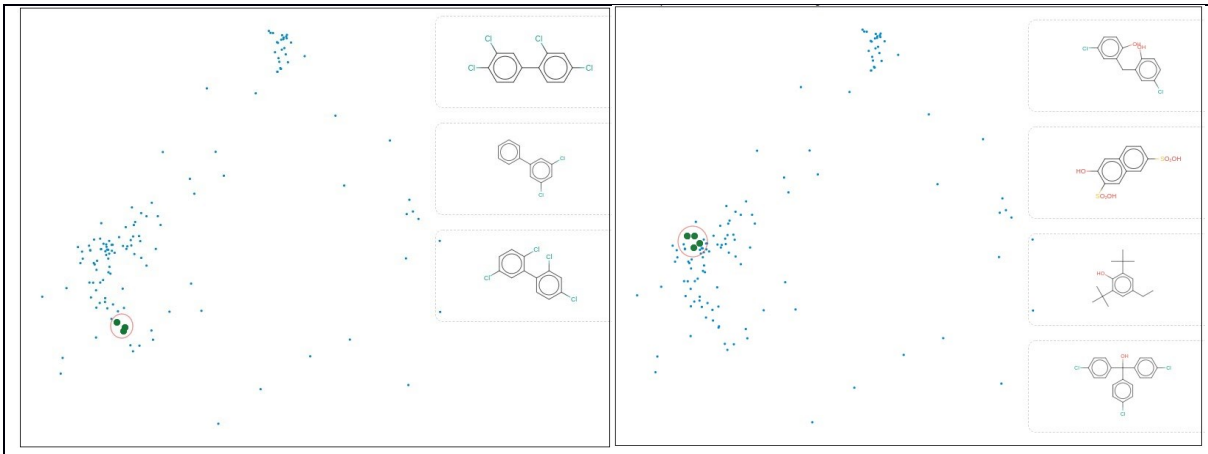
inspection of the Figure 4-4, I would say that by the same logic these models are also comparable with MT_DNN models. What led you to such a conclusion about MT_DNN performance?

The reviewer is right that we did not do the statistical tests to prove this statement. As we mention at page 68 our best MT_DNN model achieved RMSE 0.71+-0.01 where standard deviation was calculated for each endpoint over cross-validation folds and then averaged over all 29 endpoints. For all descriptor sets MT_DNN model overperofrmed all other methods by average performance. We did not perform statistical tests because of some technical limitations in multatask mode in OCHEM. Hovever, I can't disagree with the reviewer that statistical model comparison is a good practice.

(Peter Ertl): Chapters 3 and 4, quite sophisticated methods are used to predict data that show quite high experimental variation (bioconcentration and toxicity) sometimes even in range of tens of percent. It would be good to look into this, learn about variation and reliability of data and look at how this affects the reliability of predictions; it would be useful to show also some representative molecules for these datasets to see what part of chemical space we are dealing with; how the training and test sets have been selected (random or diversity-based)?

We use 5-fold cross-validation for all models in this study unless otherwise specified. We used only random splits for cross-validation folds. Albeit, for the broad chemical space like RTECS we believe that random split provides a reasonable distribution. Because the original EPA BCF dataset already was split into training/test, we calculated both 5-fold cross-validation results and the results on the test set (Table 3.1). We also included standard deviations over cross-validation folds to demonstrate the reliability of modelling. We processed BCF dataset with our parametric t-SNE model and indicated typical representatives in the figures below.





Unfortunately, we have no permission to publish all compounds from RTECS by licensing agreement. However, we presented a t-SNE projection of RTECS chemical space and a description of representative compounds for some clusters in Figure 4-2. One can see that the chemical space is quite broad and diverse. The distribution of toxicity is also given on a subfigure of Figure 4-2.

(Daniel Svozil): In Table 4-4, the Feature Net is compared with ST_DNN and MT_DNN models. Why didn't you also include RF into the comparison? According to Figure 4-4, RF is the best performing model from all single task models. Please, could you compare the performance of MT_DNN and RF and specify their advantages and disadvantages?

We did not include Random Forest experiments because Random Forest models' calculations, even for several descriptors sets over 5 fold cross-validation, required a lot of time. Our idea was to compare only neural networks in this experiment. For example, due to computational costs, we use only XGboost in similar experiments with attributed modeling. There is another problem because Feature Net is not in the standard OCHEM pipeline, we had to modify each computational method by hand to make it applicable for Feature Net. Due to some modifications, it is hard to implement Random Forest for Feature Net modeling in the current version of OCHEM. The exhaustive study of the Feature Net approach is beyond the scope of this research.

Chapter 5 (Chemical space visualization guided by deep learning):

(Peter Ertl) Chapter 5 - why the tSNE was used for initial dimensionality reduction, this method does not keep well distances between the molecules, has the other dimensionality reduction methods been considered? Would it be possible to skip the tSNE step completely and train the network to map molecules directly to the 2D space?

We used t-SNE because this method works well to visualize different types of data, and we had expertise with it. We assumed that it is reasonable for chemical structures too. This method has a clear statistical rationale and simple implementation. However, there are some issues that we mentioned in Section 5.1, and we proposed a parametric (ANN-based)

version to overcome these issues. We also compared the performance of t-SNE with PCA and MDS. Our approach for the quantitative estimation of visualization quality is described in Section 5.2. The results of this experiment are given in Table 5.1. It demonstrates that the quality of parametric t-SNE projections overperforms PCA and MDS.

It is worth mentioning that our approach is fully data-driven. It does not require any "cheminformatics" processing except descriptors calculations (moreover, I see no problem to run it directly on SMILES strings). I can assume that "chemistry-based" approaches like Molecule Cloud (10.1186/1758-2946-4-12) could provide better visualizations, but they are based on external knowledge or algorithms.

I believe that, with neural networks, one can map any representation to any other representation. *The question is: how to define a loss function for training?* We use t-SNE just because it provides a convenient way to define the loss function. It consists of three parts: a statistical definition of high-dimensional space, a statistical definition of a low-dimensional space, and a “glue” between these statistical distributions: KL divergence, that we use as a loss function in our method. Any part is exchangeable. One can use other statistical definitions or other distance functions or even other loss functions that have nothing in common with t-SNE.

(Nikolay Brilliantov) From the nature of the 2D patterns that arise in the process of mapping from the high dimensional descriptor space, one can expect that the distance matrix will be a sparse one. High Dimensional sparse matrices demonstrate many important properties. These properties of the sparse matrices may be additionally used to understand the nature and topology of the chemical space. Did the author consider such a possibility?

The original distance matrix is not sparse, but the distance between non-neighborhood compounds in a high-dimensional space would be marginal, and one can turn it to zero. However, we did not consider this possibility, because we had no ideas how the properties of high dimensional sparse matrices can help in the analysis of the internal structure of chemical space and improve the visualization of chemical compounds.

(Nikolay Brilliantov) This comment is related to the previous one: What is the motivation of the application of the loss function L in the form (5.3)? Structurally, it looks as a definition of a Complexity of a 2D Pattern with the cross-probabilities of its elements pij. Although these quantities are the elements of the distance matrix, does it have any relation to the complexity of the 2D structure which results from the mapping from the high-dimensional descriptor space? If such a connection exists, it would be worth discussing it.

This question is, in fact, not about the parametric t-SNE method but classical SNE and t-

SNE. The main idea of SNE and t-SNE is to place points on a low-dimensional plane so that the objects that are close to each other in high-dimensional space would lay nearby. The reformulating it in terms of statistical distributions allows using KL-divergence as a "distance" between distributions. One possible reason why the original t-SNE uses KL-divergence is that the loss function's gradient is quite simple. Because parametric t-SNE is the extension of classical t-SNE, it uses most of the machinery of the classical t-SNE. Regarding the question “Although these quantities are the elements of the distance matrix, does it have any relation to the complexity of the 2D structure which results from the mapping from the high-dimensional descriptor space?”, that is right that there is no guarantee that the complexity of a 2D structure reflects some high-dimensional relations of data, but in practice, we know that it provides reasonable projections. t-SNE is one of the most popular visualization methods; it proved its efficiency for many scientific and industrial applications (10.1038/s41467-019-13056-x, Laurens van der Maaten and Geoffrey Hinton Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579-2605, 10.1021/acs.jcim.8b00640).

(Nikolay Brilliantov) The application of the expression (5.1) in page 81 is not sufficiently justified/explained. The author writes: “…where the parameter □ is found by binary search to achieve the predetermined entropy of the distribution”. What is the “predetermined entropy”? What kind of entropy is used? Shannon entropy, Reni entropy or may be Tsallis entropy? At least the structure of the Eqs. (5.2) and (5.3) in page 82 leads to such a guess. The motivation of the usage of Eq. (5.1) and of its practical application should be better articulated.

We are thankful for the reviewer for this indication. First, there is a mistake in this equation: the parameter □ should be □$_i$. This parameter is tuned over batches by binary search to satisfy a predefined criterion -- perplexity. Perplexity is defined as:

$$Perp(P_i) = 2^{H(P_i)}$$

where H(Pi) is Shannon entropy:

$$H(P_i) = -\sum_{j=1}^{N} p_{ji} log_2(p_{ji})$$

This mechanism dynamically tunes and unifies the so-called “effective distance of interactions” or “ the bandwidth of the Gaussian kernels” for the points in dense regions of high-dimensional and sparse ones. One can regard it as obtaining a similar amount of information to a point from its neighbors independently of a region's density. The perplexity itself influences the performance of projection. One can see Figure 5-2, where the learning curves obtained for different perplexity values are shown.

We corrected Eq 5.1, added equations 5.2 and 5.3 to provide a detailed explanation of the perplexity and entropy.

(Daniel Svozil) Can you compare computational requirements (speed and memory demands) of the parametric t-SNE with its non-parametric version, as well as with other embedding methods such as MDS, IsoMap and UMAP? IsoMap and UMAP are also very popular, why were these two not included in the study?

When we performed experiments with parametric t-SNE, UMAP first came out, and we knew nothing about it. This algorithm became popular later. MDS and Isomap were definitely below the requirements for the processing of extra-large chemical datasets (at least their sklearn implementations), and moreover, the quality of MDS projections is lower than pTSNE, as indicated in Table 5.1 Regarding time, our current t-SNE code requires about 8 minutes per epoch for ChEMBL subset (1.1M of compounds, 1024-bit ECFP fingerprints). Memory consumption clearly depends on the batch size. Commonly we train our parametric t-SNE model for 30 epochs, however the loss saturaets quickly. Multicore t-SNE implementation (https://github.com/DmitryUlyanov/Multicore-TSNE) does not finish calculations on ChEMBL in a reasonable timeframe at all. Standard UMAP (https://github.com/lmcinnes/umap) fails on my machine with 64Gb of memory, probably due to memory overfull. The only algorithm that has succeeded with this dataset was ncvis (https://github.com/stat-ml/ncvis). It took 1h 26min 32s for ncvis to project. Basically, nowadays there are algorithms that can project millions of compounds. But the motivation for parametric t-SNE was not only the processing of extra-large datasets but the ability to apply it to new compounds. Without such a feature the generation of compounds described in Chapter 6 would be impossible.

(Daniel Svozil)I (If I understand it well, the cost function (the Kullback-Leibler divergence in your case) can be used as the measure of map quality (i.e., the map with lower cost function can be regarded as better). This is useful when you want to compare different maps generated with different algorithm settings. Please, could you comment on how such defined quality relates to the real-use quality of the map? By this I mean if it can, e.g., happen, that maps with lower cost functions may reveal relationships not observed in "better" maps.

The problem of quantitative estimation of visualization quality is a crucial part of this research. We proposed a method for the quantitative estimation of it. We describe it in Section 5.2 In short, we train a simple classifier for a set of projected bioactive molecules. This classifier has only 2D coordinates on its input and predicts the bioactive class. If the projection is reasonable the classifier can follow the chemical paradigm that "similar compounds have similar properties" and (somehow) succeed in prediction. In case of unreasonable projection, similar compounds are not grouped together and the classifier has no chances. Comparing the performance of classifiers, we can rank the reasonability of our projection. We believe that just comparing KL-divergences is not a good idea, and the reviewer's concern that "maps with lower cost functions may reveal relationships not

observed in “better” maps” is sensible.

Chapter 6 (Optimized molecular grammars for structures generation):

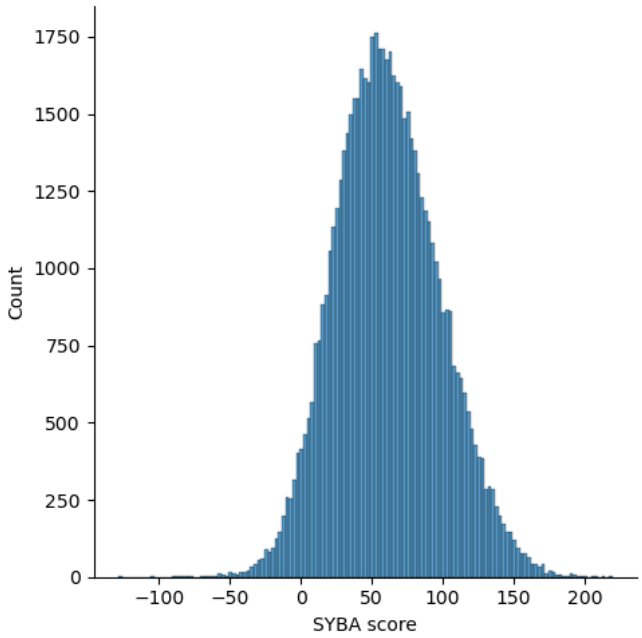
(Peter Ertl) In the overview of using SMILES in deep learning, discuss also variant of SMILES designed specifically for deep learning: O’Boyle N, Dalke A. DeepSMILES: An Adaptation of SMILES for Use in MachineLearning of Chemical Structures. https://europepmc.org/article/ppr/ppr56249

I added the description of DeepSMILES to the introduction to Chapter 6.

(Nikolay Brilliantov) Is the ability of Legogram to produce “absolutely correct” molecules is determined by the “block” nature of this approach, that is, by the fact that it operates with existing chemical groups, which chemical combination produces correct molecules with much high probability that chemical combination of single atoms?

The ability of Legogram to produce chemically valid molecules is not based on probabilities. Legogram builds molecules from “blocks” but blocks can represent just single atoms. Moreover, the encoding process leads to blocks that are either atoms or rings. The ability of Legogram to produce “absolutely correct” molecules based on the fact that for each block we calculate its connectivity patterns and Legogram combines together only blocks that preserve chemical validity. For example, Legogram just can’t connect a halogen atom to a block that represents “a carbon atom with a double bond” but can do it for “a carbon atom with a single bond”.

(Daniel Svozil) I wonder if compounds generated by the legogram are synthetically accessible. How is this ensured?
To elaborate this question we used the SYBA package that scores compounds by it’s synthetic accessibility. We calculated scores for all compounds generated by *Agent* network during QED optimization. One can see that the majority of compounds have SYBA score above zero, that means that are regarded as easily synthesible. We added a subsection “Synthetic accessibility” to Chapter 6.



(Daniel Svozil) One of the most important quantities generated compounds to be optimized for is the biological activity against the selected biological target. How would you bias your generator to yield compounds with higher probability being biologically active?

Our approach is based on a well-studied REINVENT method (but they use SMILES, while we use Legogram as a generative backbone), This method allows to bias a generator to any property, including bioactivity. The authors of REINVENT demonstrated it experimentally for DRD2 inhibitors. We believe that one can use our method too for the generation of bioactive compounds. We added the discussion to Section 6

(Peter Ertl) In the overview of using SMILES in deep learning, discuss also variant of SMILES designed specifically for deep learning: O'Boyle N, Dalke A. DeepSMILES: An Adaptation of SMILES for Use in MachineLearning of Chemical Structures. https://europepmc.org/article/ppr/ppr56249

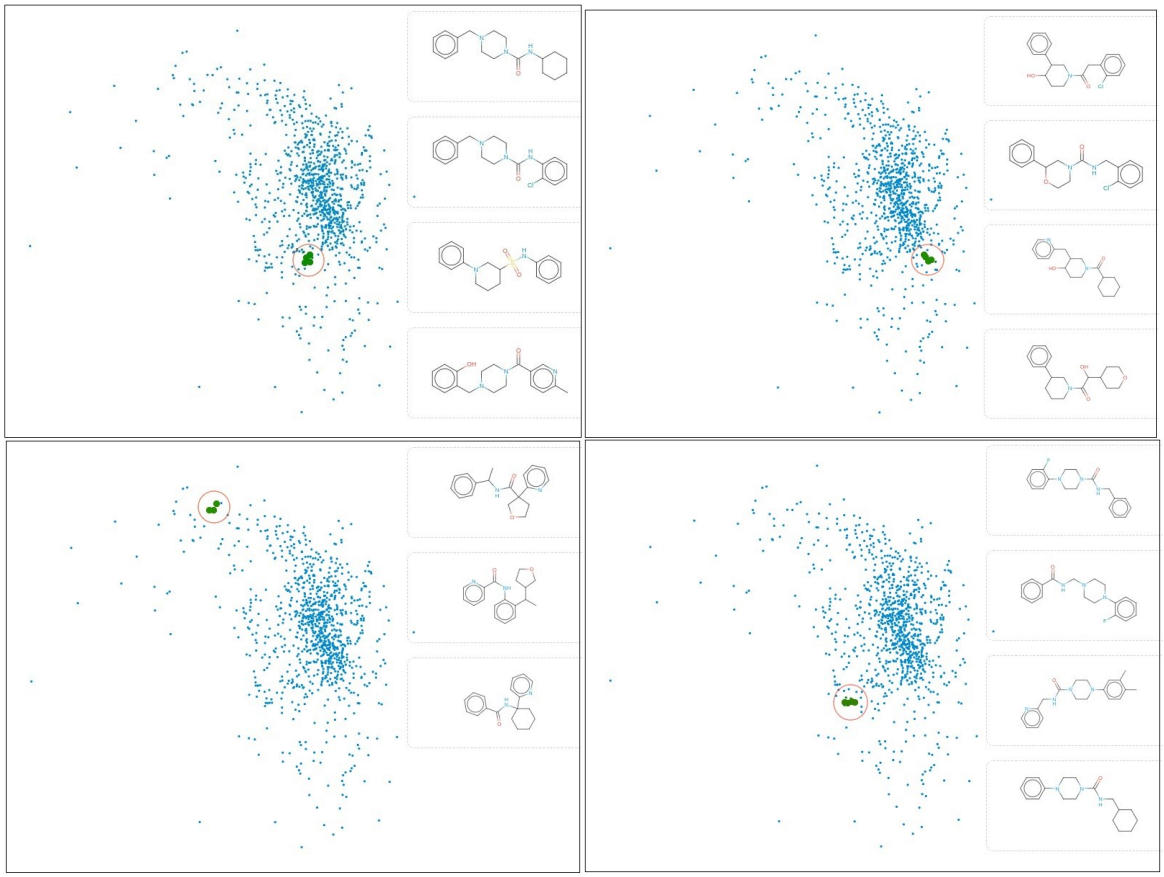
I added the description of DeepSMILES to the introduction of Chapter 6.

(Peter Ertl) Chapter 6 - it would be interesting to see some simple statistics of properties of generated molecules, for example % of purely aromatic or aliphatic molecules, how many of them contain rings, are macrocycles, how many are pure hydrocarbons etc.

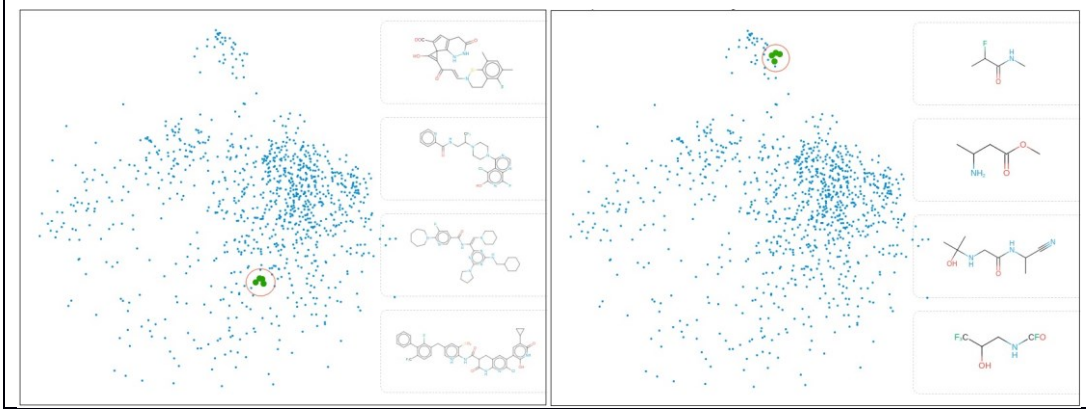
We provided a parametric t-SNE visualization of chemical space for 1000 molecules with the best QED values and 1000 molecules with the worst QED with typical molecules in some clusters. One can see that for top QED set there are no "simple" molecules at all, however the

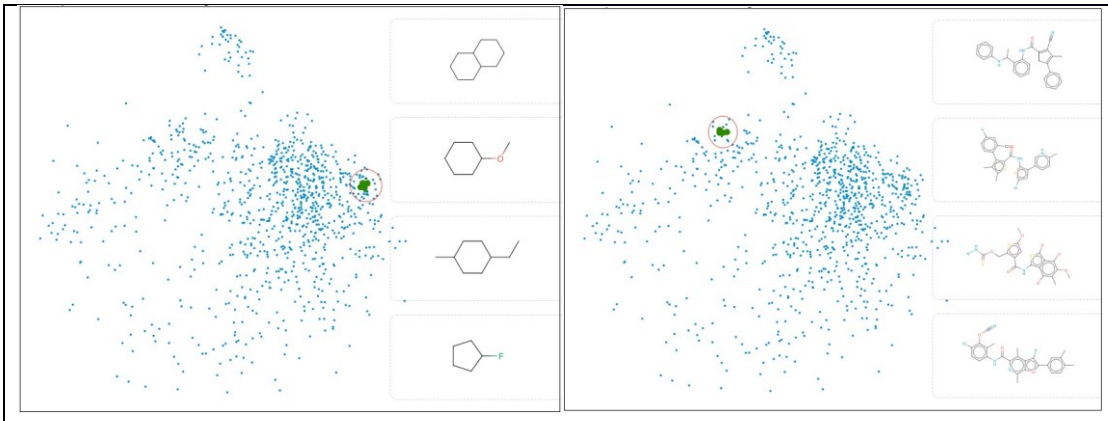
set is quite diverse. For worst QED we have some “simple” molecules (methane, short hydrocarbons etc) and some quite complicated ones. The chemical space looks more diverse, which is not surprising. So, our generative apprach can cover broad chemical space.

QED high:



QED low:





Minor comments and corrections:

(Maxim Panov) [Section 2.3] It appears that square is missing in both starting formulas of the section for the part of the formula to the left of equality sign. Also, the variance term should include only Var[\hat{f}] as Var(y) is exactly sigma^2 (next summand).

I fixed it.

(Maxim Panov) [Section 2.4] \rho_\gamma is not defined.

It was defined at page 37, however I added the description of \rho_\gamma and \rho_\gamma(\rvec) just after the formula 2.33

(Peter Ertl): Is the correlation shown in Figure 3.5 for training or test set?

This figure is for the test set.

(Maxim Panov) [The sentence before formula 2.3] Should be WiEner index

I fixed it

(Maxim Panov) [Formulas 2.3 and 2.4] The same letter is used for degrees and distances, might be confusing.

I changed it to deg(v)

(Maxim Panov) [Explanation of formula 2.5] Should be F_i is a molecular fragment.

I fixed it

(Maxim Panov) [Formula 2.6 and everywhere else] If upper index for summation is provided

then lower index also should be present, i.e. sum_{i = 1}^N.

I fixed it in Eq. 2.6, Eq. 2.7 Eq. 2.24 Eq. 2.25 Eq. 2.26

(Maxim Panov) [Formula 2.8] \theta is not introduced.

I added "$ \theta $ -- is a parameter set"

(Maxim Panov) [Formula 2.2] It is not clear with respect to what expectation is taken. I also can only deduce that it is expectation as this symbol is not explained in the text.

I added an additional explanation to Formula 2.2

(Daniel Svozil) Table 5.1. The horizontal blocks (i.e., kNN, SVM, XGBoost, RF) are not described which makes the understanding more difficult.

I fixed it

(Daniel Svozil) Page 110-111. The last paragraph on the page 110 does not, apparently, belong here. It looks like the response to the reviewer comment.

The paragraph should be here, because it answers an important question raised directly in the text. However, I reformulated this question as: "does our model follow a global chemical space"

(Daniel Svozil) Chapter 2.6 Bioconcentration factor paragraph does not contain the description of data sets used, while the next two paragraphs (Acute animal toxicity and Bioactivity datasets) do.

I changed it by fully dismantling the section 2.6 and move the information from this section into corresponding "Materials and Methods" sections

(Nikolay Brilliantov) Page 31 and 38: Please read carefully to remove typos

I fixed some typos on these pages.

(Nikolay Brilliantov) Page 25: I believe that there is a discrepancy in the notations – the author should use either pi or Δ pi in Eq. (2.6) and the text below the equation.

I fixed it

(Nikolay Brilliantov) Page 27: after Eq. (2.8) "We can thing"..... please correct

I fixed it

(Nikolay Brilliantov) Page 26: I expect that the upper limits in the summation in Eq. (2.7) should be L and R in the first and second term respectively

I fixed it

(Maxim Panov) [Formula 2.9] x_i is not introduced, it is not clear how y_t depends on x_i.

(Maxim Panov) [Formula 2.10] It is not clear for what reason we have some constant C here.

(Maxim Panov) [Formula 2.12] Again, index i should be somewhere in y_t.

(Maxim Panov) [Formula 2.15] Some w_i appear, again not clear who are they

(Maxim Panov) [Formula 2.18] There is a problem with indices.

(Nikolay Brilliantov) Page 28: In Eqs. (2.16) and (2.17) are the quantities wij and the same quantities with a hat the same?

I revised the structure of this section and removed the detailed description of XGBoost because it was too detailed compared to other methods, and not so relevant to overall structure.

(Peter Ertl) Some small formal changes: literature: change Robert C. Glem to Robert C. Glen, Figure 6.6 - is not clear, atom labels not visible, redraw more clearly

I fixed it, it was an incorrect record in PubMed (https://pubmed.ncbi.nlm.nih.gov/16523386/)
I redrew Figure 6.6, however leave only 4 representative rules.

(Nikolay Brilliantov) Page 25: The letters in the diagram are hardly visible, the picture is fuzzy, a better quality figure is required

I recreated this figure

(Daniel Svozil) Page 69. The number of records for each endpoint is missing in Table 4.1, in contrary to what is stated.

Table 4.1 is the chemical descriptors in OCHEM. The number of compounds for each endpoint is given in Table 8.5: "Endpoints extracted from RTECS dataset in Supplementary" Material"

(Nikolay Brilliantov) Page 20: the awkward expression: ..."we are use to using".... please correct

Now it is “Commonly, molecular descriptors are inputs for machine learning models”