3D matters! 3D-RISM and 3D convolutional neural network for accurate bioaccumulation prediction

3D matters! 3D-RISM and 3D convolutional neural network for accurate bioaccumulation prediction

Sergey Sosnin[1], **Maksim Misin**[2], **David S. Palmer**[3], **Maxim V. Fedorov**[1],[4]

[1]Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow 143026, Russia
[2]Institute of Chemistry, University of Tartu, Ravila 14a, Tartu 50411, Estonia
[3]Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, UK
[4]Department of Physics, University of Strathclyde, Scottish Universities Physics Alliance (SUPA), Glasgow, UK

E-mail: m.fedorov@skoltech.ru

Abstract. In this work, we present a new method for predicting complex physical-chemical properties of organic molecules. The approach utilizes 3D convolutional neural network (ActivNet4) that uses solvent spatial distributions around solutes as input. These spatial distributions are obtained by a molecular theory called three-dimensional reference interaction site model (3D-RISM). We have shown that the method allows one to achieve a good accuracy of prediction of bioconcentration factor (BCF) which is difficult to predict by direct application of methods of molecular theory or simulations. Our research demonstrates that combination of molecular theories with modern machine learning approaches can be effectively used for predicting properties that are otherwise inaccessible to purely theory-based models.

Submitted to: *J. Phys.: Condens. Matter*

1. Introduction

Molecular theories such as three-dimensional reference interaction site model (3D-RISM) [1, 2, 3], ER-theory[4] or molecular density functional theory (MDFT) [5, 6, 7] rely on approximations derived from rigorous statistical mechanics to estimate the equilibrium distribution of solvent around solvated molecules. In turn, these distributions can be related to many physical-chemical properties of a solvated molecular system[8, 9]. Examples of such properties include solvation free energy[10, 11, 12], partial molar volume [3, 13], salting-out constants [14] and binding free energies [15, 16, 17]. However, using a purely theoretical approach, it is difficult to relate these distributions to the substance's biological effects which are a result of a large number of complex interrelated phenomena, such as toxicity or bioaccumulation.

The above does not mean that the solvation structure is not useful for the understanding of the influence of chemical compounds on the living organisms. On the contrary, the information encoded in the solvation shell can be used to understand whether a given compound is hydrophobic or hydrophilic [18] which in turn can provide a reasonable guess whether it will be able to pass certain membrane channels [19]. In case of a solution which contains ions, the solvation structure can provide and estimation the solute affinity towards them [14]. All this information is directly *related* to compound's biological effects but can not be expressed *explicitly* using equations. On the other hand, machine learning methods are usually quite good at finding and quantifying such 'hidden' relations [20, 21, 22].

In this article, we utilize a 3D convolutional neural network (CNN) to develop a prediction model which can estimate the bioaccumulation propensity of a compound characterised by the bioconcentration factor (BCF) for a number of different organic molecules. As an input, we use three-dimensional distributions of water around these molecules, obtained by 3D-RISM with Kovalenko-Hirata closure (KH) [23]. Artificial neural networks (ANNs) have been previously used for predicting biological effects of organic molecules [20, 21, 22]. However, they were combined with a very broad set of descriptors that have diverse physical meanings. Here we focus on a single descriptor; solvation shell structure in an attempt to show that this can be a universal descriptor for prediction of properties of molecules which are difficult to formalise by a theory. To determine whether the CNN-based machine learning setup is necessary, we also tested linear and Extreme Gradient Boosting (XGBoost) models and compared them with the 3D CNN approach.

2. Theoretical Background

2.1. 3D Reference Interaction Site Model (3D-RISM)

Calculation of an equilibrium distribution of solvent around an arbitrary molecule is a challenging problem in computational molecular science [3]. It can be done by molecular dynamics simulations, but extremely long simulation times are needed to obtain smooth solvent distributions [24].Theoretical methods of statistical mechanics like MDFT method can be applied to this problem as well as the 3D-RISM theory [1, 3, 23, 25, 26, 27] which is used in this work.

As a result of the 3D-RISM calculation, one obtains *density distribution functions* (local densities) $\rho_\gamma(\boldsymbol{r})$ of every of the solvent sites γ around the solute molecule. We used a 3-point model of water (SPC/E) meaning that the calculation produces density distributions for both oxygen and hydrogen atoms. These density distributions can be regarded as a variant of molecular fields. Notice that the densities obtained from RISM calculations are not exact [2, 13], but can be successfully used to predict a variety of physical properties using either empirical or semi-empirical corrections [12, 14, 28, 29, 29] or QSPR approaches [30].

The 3D-RISM main equation can be written as [2, 13]:

$$h_\gamma(\boldsymbol{r}) = \sum_{\alpha=1}^{n_s} (\chi_{\alpha\gamma} * c_\alpha)(\boldsymbol{r}),$$

where $*$ denotes convolution, n_s stands for the number of solvent sites, and $h\gamma(\boldsymbol{r}) = \rho_\gamma(\boldsymbol{r})/\rho_\gamma - 1$, is usually referred to as the total correlation function. One should note that $\chi_{\gamma,\alpha}$ is obtained from a homogeneous solvent, while h_γ, c_α are solute-solvent correlation functions that describe a system with a fixed solute molecule, surrounded by solvent. $c(\boldsymbol{r})$ is a direct correlation function[27]. Finally, $\chi_{\alpha\gamma}(r)$ is a site-site susceptibility function that can be obtained from a bulk solvent radial distribution functions. More conveniently, $\chi_{\alpha\gamma}$ can be calculated from a separate 1D-RISM calculation[3, 31].

The above equation is coupled with a separate closure relation that provides another connection between $h_\gamma(\boldsymbol{r})$ and $c_\alpha(\boldsymbol{r})$. In this work we used standard, computationally robust (KH) closure [32]:

$$h_\gamma(\boldsymbol{r}) + 1 = \begin{cases} \exp\left[-\beta u_\gamma(\boldsymbol{r}) + h_\gamma(\boldsymbol{r}) - c_\gamma(\boldsymbol{r})\right], & \text{if } h(\boldsymbol{r}) \leq 0; \\ 1 - \beta u_\gamma(\boldsymbol{r}) + h_\gamma(\boldsymbol{r}) - c_\gamma(\boldsymbol{r}), & \text{if } h(\boldsymbol{r}) > 0; \end{cases}$$

where $\beta = 1/(kT)$ and $u_\gamma(\boldsymbol{r})$ is a potential energy between the solvent site γ and the solute molecule. Together the above systems of equations are usually iteratively solved until both $h_\gamma(\boldsymbol{r})$ and $c_\alpha(\boldsymbol{r})$ achieve a predefined convergence criteria.

2.2. Bioconcentration factor

In this work, we built a model for predicting the BCF (more specifically, we predicted its decimal logarithm $\log_{10}$BCF). This factor is the ratio between the concentration of an organic compound in biota and in water:[33]

$$BCF = \frac{C_{biota}(\text{compound})}{C_{water}(\text{compound})}, \tag{1}$$

where C represents concentration. It should be noted that BCF is regarded as a consolidated property of a chemical compound, due to this reason the definition involves some common concepts like "biota" and "stationary concentration in vivo". However, there is OECD 305 guideline[34] which provides the basic requirements for the methods that should be used for bioconcentration factor estimation to obtain high quality and comparable data. The typical way to estimate BCF is a measurement of the concentration of a compound in fishes and in the water after reach of stationarity of concentrations, usually by exposing the chemical during the pre-defined long time period. Strictly speaking, there are many types of BCF which definitions depends on the concentrations of compounds, species of animals, times of expositions, and other factors of the experiments, but due to the generality of BCF nature, the merging of BCF values of different experiments (even for different fish species) is possible. This factor is an important parameter for estimating the potential danger of an organic compound. It is one of the parameters that determine the labeling of the compound under *Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)* program. The ability of a compound to penetrate and remain in an organism may influence the toxicity and mutagenicity, and so may reveal potential environmental risks. Generally, if a chemical has BCF value of more than 5000 (or $\log_{10}\mathrm{BCF} > 3.67$), it is regarded as potentially dangerous. There are several methods to measure and estimate the confidence of the BCF data, described in details in ref. 35. It should be emphasized that determining of BCF in in-vivo experiments is a very expensive procedure.

Over the years, several models for the BCF prediction have been published. Arnot and Gobas have proposed a linear model that predicted BCF as a function of the uptake and elimination of an organic compound by an aquatic organism. Since BCF is related to logP and water solubility[35], some authors proposed models that utilised these descriptors [36]. These linear models work satisfactory only for moderately hydrophobic compounds, but fail to address strongly hydrophobic chemicals[37]. Additionally, *LogP* and solubility must be measured separately and this may be problematic. Another notable model has been produced by Zhao et al.[38] using a hybrid of a number of machine learning methods. Their model managed to produce an impressive accuracy ($R^2 = 0.8, \mathrm{RMSE} = 0.59$), albeit on a somewhat curated dataset.

To conclude the above, modeling of the bioconcentration factor is an important research area due to the difficulties associated with its experimental evaluation and importance of such models for regulatory purposes.

3. Methods and Materials

Database We used the dataset collected by USA Environmental Protection Agency for their T.E.S.T. QSAR platform for risk estimation[39]. US EPA collected the database from several sources[38, 40, 41]. This dataset contains BCF measured values for several fish species: european carp and salmonids. As it has been discussed above combination of BCF values from cross-species experiments is allowable. We did not do any changes (modifications, additions, filtration) in the dataset. This dataset has been split into training and test subsets in the same manner as it was done by US EPA, and statistical values on the test set are published. We used them as a baseline for our model. There are 541 molecules in the training set and 135 molecules in the test set.

We used RDKit[42], an open-source cheminformatics toolkit, to perform basic molecular routines and to estimate the geometries of molecules.

Conformers Generations For deep neural networks, high amount of diverse data is a key factor to success. Our approach to conformer generation and selection is similar to the method from the article [43] and is briefly described below.

At the first stage of the algorithm, we generate a number of conformers by rotating the bonds of a molecule in stochastic manner. This is followed by an energy minimization step, consisting of 5000 iterations and performed using the universal force field (UFF) [44]. Subsequently, the set was pruned such that only conformers with mutual RMSD (computed on the heavy atoms) more than 0.5 Å have been kept. If the number of conformers exceeds the pre-defined limit, then the post-processing procedure from paper [43] is performed (we discuss this procedure in more details in the Supporting Information). We note that the prediction output for every molecule is an average over the whole ensemble of corresponding conformers. We believe that this procedure can also mitigate potential issues related with rotations of molecules.

3D-RISM Calculations We used AmberTools16[45] package to calculate the partial charges of each molecule using AM1-BCC[46] semi-empirical model. At this stage, for some molecules the calculations have not converged, and these molecules were eliminated. These partial charges were used for further 3D-RISM calculations. All 3D-RISM computations were performed using *rism3d.snglpnt* program[23, 24, 26, 27] from AmberTools16[45] package. Site-site susceptibility functions of bulk water $\chi_{\alpha\gamma}(\boldsymbol{r})$ were calculated using DRISM method by *drism* program from the same package. The water temperature was set to 298 K. For 3D-RISM we used a 35 Å × 35 Å × 35 Å grid with 0.5 Å step size. The resulting oxygen and hydrogen density distributions were saved as HDF5[47] binary files. We ran a separate 3D-RISM calculation for each conformer. If more than 50% of 3D-RISM calculations did not converge, such molecule was eliminated from the dataset.

3D Convolutional Neural Networks Modeling Procedure We used framework chainer[48] to build networks for processing 3D data. The architecture of the network is schematically presented in figure 2 (a more standard representation is provided in table S1 in the Supporting Information). This architecture was optimal in terms of speed and the quality of the training models. This model has been called *ActivNet4*, with four indicating the number of convolutional layers used. A *pooling* layer is introduced in the structure of the CNNs which reduces to a minimum the potential effects of translation, rotation and shifting of molecules on the final output of the algorithm.

We trained this network using both oxygen and hydrogen density distributions, obtained from 3D-RISM calculations.

Parametric Rectified Linear Units[50] were used as activation functions for the model since they showed small improvement in the prediction quality, although, it is possible to replace them with the commonly used *relu* activation function without a noticeable lack of performance. To train *ActivNet4*, we experimented with several

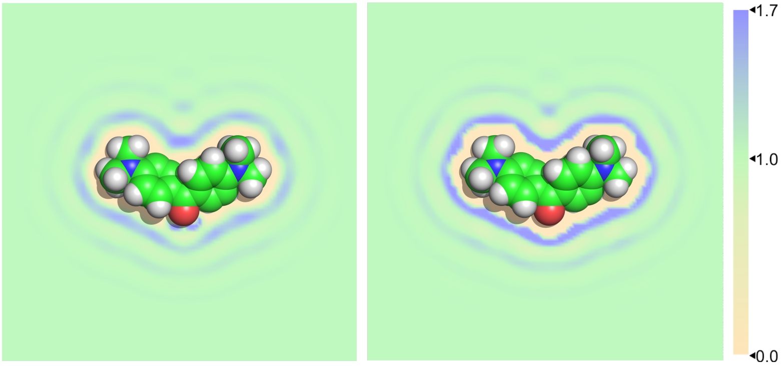

Figure 1: An example of the visualization of the scalar fields for a molecule as 2D slices taken by the principal axis (Left – a visualization of hydrogens density. Right – a visualization of oxygen density. Light yellow color – lower values, pale green color – bulk values, blue color – higher values))

optimizers: Stochastic gradient descent with momentum, *Adam*[51], *RMSprop*[52], and *SMORMS3*[53]. The best and the most stable convergence has been provided by *SMORMS3* method. *RMSprop* and *Adam* have a good convergence ability, but the training process was less stable. Stochastic gradient descent has converged noticeably more slowly for the network. The parameters of the optimisers can be found in the Supporting Information.

The training and test procedures slightly differed. At the training stage, each conformer of the molecule has been regarded independently from the other conformers. At the test stage, the prediction value for each conformer of the molecule has been calculated and the final result was the mean value for all conformers of the molecule. The performance of the model was estimated on the same test set that has been used in the original work to compare our model with the baseline. Additionally, we used a 5-fold cross-validation (CV) technique for the whole dataset to measure the quality of the model in a more reliable way. The Neural networks have been trained using Nvidia K80 graphics cards and Nvidia GTX 1080 cards. Training of one model requires approximately 5 hours on Nvidia GTX 1080 and up to 4 times longer on Nvidia K80 graphics cards.

Extreme Gradient Boosting modeling (3D Fields) To compare our 3D convolutional network with other machine learning approaches we built a model using Extreme Gradient Boosting (XGBoost implementation[54]) algorithm. This method has been proposed for use in cheminformatics[55] and can process very large datasets. In this experiment, initially, we had to decrease the volume of each 3D cube from *70x70x70* to *17x17x17* by performing the average pooling operation with a kernel *(4,4,4)*. Then, both oxygen and hydrogen channels have been flattened and stacked forming a vector of *9826* values. These vectors served as the inputs for XGBoost algorithm. The application of the method to the test set has been performed in the same manner as in the neural network experiment. We used the maximal number of trees = 100 and maximal depth

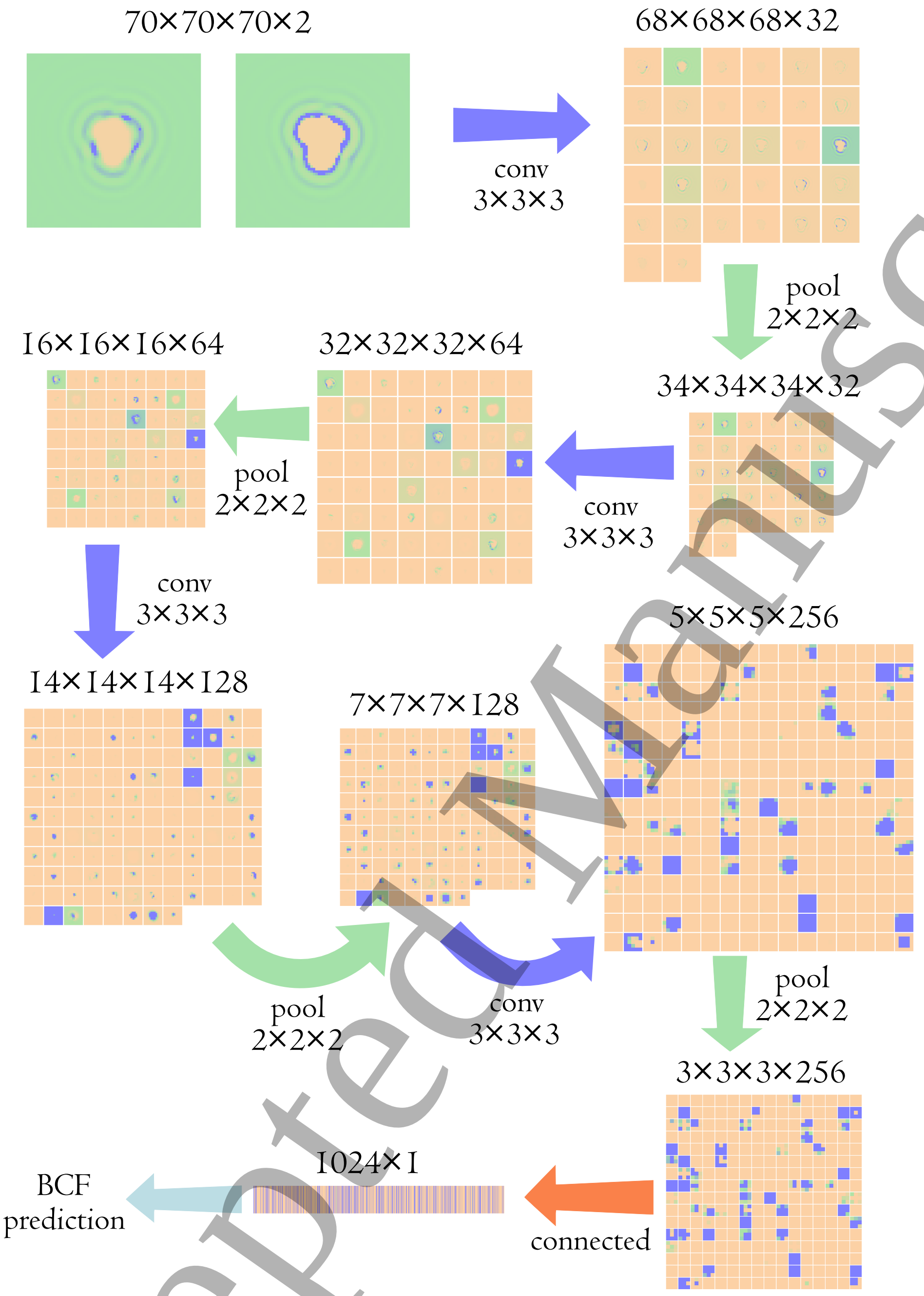

Figure 2: A schematic representation of *ActivNet4* architecture with visualized 2D slices of feature maps on a trained network. Feature maps are colored using the same color scheme as in figure 1. Blue arrows labeled conv N × N × N denote a 3D convolution layer, green arrows labeled pool N × N × N denote 3D max-pulling layer, and red arrow labeled "connected" denotes a fully-connected layer. The figure is based on figure 4 from Ref. 49

of each tree = 6 to train the models, the other parameters have been set to default.

Graph Convolution modeling It was shown recently that in some cases graph convolution methods can overperform traditional QSAR/QSPR approaches which are based on the molecular descriptors[56, 57]. We used DeepChem[58] framework included in Online Chemical Modeling Environment[59] to build graph convolutions models. For graph convolution model we used the hyperparamethers: epochs 100, learning rate 0.001, dropout rate 0.25, dense layer size 128 neurons, the size of convolution layers was (64,64) the other parameters have been set to default.

Linear model Finally, we also built a linear model for BCF prediction using the following relation:

$$\log_{10} \mathrm{BCF} = a_1 \Delta G + a_2 \bar{V} + a_3, \qquad (2)$$

where ΔG is molecule's hydration free energy, obtained with 3D-RISM PC+ method [13, 60], $\bar{V}$ is partial molar volume, and a_i are parameters adjusted in the process of regression. The optimal results were obtained with $a_1 = 0.10634 \frac{mol}{kkal}$, $a_2 = 0.003\,57\,\text{Å}^{-3}$, and $a_3 = 1.64677$.

4. Results and Discussion

Our main goal was to evaluate whether it is possible to predict biological property using a combination of solvation structure and machine learning. For this, we took EPA database which has 676 molecules with known BCF. 670 molecules were successfully processed, while 6 molecules failed at the partial charges calculations stage or at the 3D RISM stage. The database were split into a training (537 molecules) and test (133 molecules) sets. For each molecule we then generated a diverse set of conformers, using a procedure described earlier. The distribution of a number of conformers for both training and test sets is shown in figure 3. As one can see, about a quarter of the molecules in the training and test sets have less than 10 conformers (quite inflexible), while the remaining molecules are highly flexible with 90-100 conformers. The distribution of the conformers is similar in both sets.

The main result of the paper is presented in Table 1. ActivNet4 model has been capable of achieving accuracy comparable to the "consensus" model provided by the US EPA[39]. This result is noteworthy due to the fact that our model was based only on the 3D distribution of water molecules while the EPA's models used a large set of descriptors of varying nature. The comparison of the two models demonstrates that the analysis of the solvent density distribution using neural networks may be useful for predicting biological properties. Surprisingly, graph convolution model showed notably worse result than baseline model, this effect can be related with relatively small dataset.

One of the problems in our approach is related to the complexity of its set up. To validate the necessity of using 3D convolutional neural networks we created Extreme Gradient Boosting (XGBoost) and linear models on the basis of 3D-RISM results. Both alternatives demonstrated poorer accuracy compared to the original, indicating that

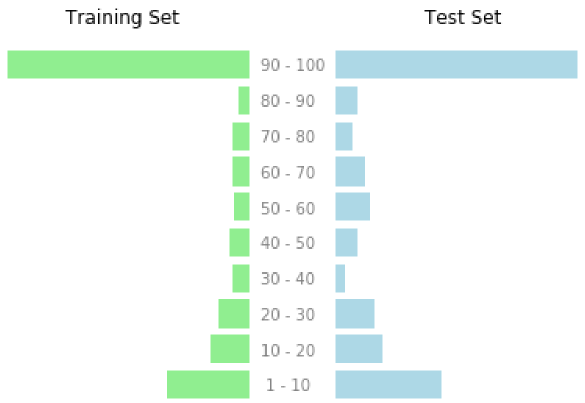

Figure 3: The distributions of the number of conformers for each molecule in the training and test sets

Table 1: Accuracies of $\log_{10}$ BCF predictions by different models. RMSE stands for root mean square error, MAE stands for mean absolute error and R denotes Pearson's correlation coefficient. For cross-validadated models the standard deviations have been calculated.

Model		**RMSE**	**MAE**	**R²**
US EPA (baseline)	consensus model	0.66	0.51	0.76
	single model	0.68	0.64	0.74
ActivNet4 (3D data)	training/test	0.66	**0.48**	**0.77**
	5-fold CV	0.65 ±0.04	0.48 ±0.01	0.77 ±0.03
XGBoost (3D data)	training/test	0.85	0.70	0.61
	5-fold CV	0.91 ±0.02	0.72 ±0.02	0.54 ±0.04
Graph Convolution	training/test	0.85	0.67	0.61
	5-fold CV	0.84 ±0.03	0.66 ±0.02	0.62 ±0.02
Linear Regression (ΔG and $\bar{V}$)	training/test	1.11	0.92	0.32

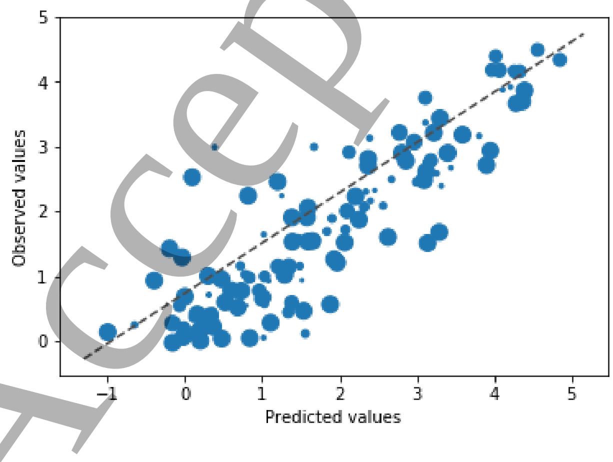

Figure 4: Correlation between observed and predicted values of $\log_{10}$ BCF. The size of the marker depends on the number of conformers of the molecule.

deep learning is necessary to achieve accurate results. To address the difficulty of the set up we created a convenient script to simplify the procedure[61].

Another bottleneck of the proposed techniques is the size of the 3D fields. For a 70 × 70 × 70 point 3D grid one has to spend a minimum of $4\,\mathrm{B}\cdot 70^3 = 1\,372\,000\,\mathrm{B}$ (1.31 MB) to store it. In the case of an n-site model of the solvent coupled with an m conformer representation of the solute we arrive to $4\,\mathrm{B}\cdot 70^3\cdot m\cdot n$ bytes necessary for each molecule.

5. Conclusions

The aim of the paper was to demonstrate that average solvent distribution in the neighbourhood of solutes can be combined with machine learning algorithms to predict properties that do not necessarily follow from the theory alone. In order to achieve it we decided to focus on predicting the bioaccumulation factor using an approximate solvent density obtained using 3D-RISM method of integral equation theories. After training, the *ActivNet4* (4-layer convolutional neural network) managed to predict $\log_{10}$BCF from water density distribution with RMSE=0.66. Although the model used relatively simple 3D descriptors, it is capable achieve prediction accuracies that are comparable to the state of the art models.

Despite the successful first results, the presented method requires a number of further developments. The first task that authors are working on right now is the application of the method to other molecular properties that are difficult to measure. Additionally, it is useful to explore possibilities of integrating solvation shell calculations and training steps to avoid storage limitations. Finally, given a clear physical meaning of the descriptors used in this study, it would be useful to explore precisely which molecular features significantly affect BCF. We hope to answer these and other questions in a follow-up article.

The source code for the 3D fields generation is located on Zenodo doi: 10.5281/zenodo.835526 and GitHub https://github.com/sergsb/clever. It is distributed under Apache License 2.0.

6. Acknowledgements

Some part of HPC calculations have been performed on Skoltech Pardus cluster. *Instant JChem* was used for structure database management, *Standardizer* was used for structure canonicalization and transformation[62]. This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", http://ckp.nrcki.ru/. Authors are thankful to Yermek Karpushev, Evgeny Burnaev and Alexey Zaytsev from CDISE Centre of Skolkovo Institute of Science and Technologies for fruitful discussions.

References

[1] Beglov D and Roux B 1997 *J. Phys. Chem.* **101** 7821–7826

[2] Hirata F 2003 *Molecular Theory of Solvation* (New York: Kluwer Academic Publishers) URL `http://www.springer.com/chemistry/electrochemistry/book/978-1-4020-1562-5`

[3] Ratkova E L, Palmer D S and Fedorov M V 2015 *Chem. Rev.* **115** 6312–6356 ISSN 0009-2665 URL `http://dx.doi.org/10.1021/cr5000283`

[4] Matubayasi N and Nakahara M 2000 *The Journal of Chemical Physics* **113** 6070–6081 ISSN 0021-9606 URL `http://aip.scitation.org/doi/abs/10.1063/1.1309013`

[5] Jeanmairet G, Levesque M, Vuilleumier R and Borgis D 2013 *J. Phys. Chem. Lett.* **4** 619–624 ISSN 1948-7185 URL `http://dx.doi.org/10.1021/jz301956b`

[6] Ramirez R and Borgis D 2005 *J. Phys. Chem. B* **109** 6754–6763 ISSN 1520-6106 URL `http://dx.doi.org/10.1021/jp045453v`

[7] Gendre L, Ramirez R and Borgis D 2009 *Chemical Physics Letters* **474** 366–370 ISSN 0009-2614 URL `http://www.sciencedirect.com/science/article/pii/S0009261409004175`

[8] Hansen J P and McDonald I R 2013 *Theory of Simple Liquids, Fourth Edition: with Applications to Soft Matter* 4th ed (Amstersdam: Academic Press) ISBN 978-0-12-387032-2

[9] Ben-Naim A 2006 *Molecular Theory of Solutions* (Oxford: OUP) ISBN 978-0-19-929969-0

[10] Du Q H, Beglov D and Roux B 2000 *J. Phys. Chem. B* **104** 796–805

[11] Palmer D S, Frolov A I, Ratkova E L and Fedorov M V 2010 *J. Phys.: Condens. Matter* **22** 492101 ISSN 0953-8984 URL `http://iopscience.iop.org/0953-8984/22/49/492101`

[12] Misin M, Fedorov M V and Palmer D S 2015 *J. Chem. Phys.* **142** 091105 ISSN 0021-9606, 1089-7690 URL `http://scitation.aip.org/content/aip/journal/jcp/142/9/10.1063/1.4914315`

[13] Misin M 2017 arXiv: 1704.05246 URL `http://arxiv.org/abs/1704.05246`

[14] Misin M, Vainikka P A, Fedorov M V and Palmer D S 2016 *The Journal of Chemical Physics* **145** 194501

[15] Genheden S, Luchko T, Gusarov S, Kovalenko A and Ryde U 2010 *J. Phys. Chem. B* **114** 8505–8516 ISSN 1520-6106 URL `http://dx.doi.org/10.1021/jp101461s`

[16] Gussregen S, Matter H, Hessler G, Lionta E, Heil J and Kast S M 2017 *J. Chem. Inf. Model.* **57** 1652–1666 ISSN 1549-9596 URL `http://dx.doi.org/10.1021/acs.jcim.6b00765`

[17] Sugita M and Hirata F 2016 *J. Phys.: Condens. Matter* **28** 384002 ISSN 0953-8984 URL `http://stacks.iop.org/0953-8984/28/i=38/a=384002`

[18] Lum K, Chandler D and Weeks J D 1999 *J. Phys. Chem. B* **103** 4570–4577 ISSN 1520-6106 URL http://dx.doi.org/10.1021/jp984327m

[19] Roux B and Karplus M 1991 *Biophysical Journal* **59** 961–981 ISSN 0006-3495 URL http://www.sciencedirect.com/science/article/pii/S0006349591823116

[20] Myint K Z, Wang L, Tong Q and Xie X Q 2012 *Molecular Pharmaceutics* **9** 2912–2923 pMID: 22937990 (*Preprint* http://dx.doi.org/10.1021/mp300237z) URL http://dx.doi.org/10.1021/mp300237z

[21] Ajmani S and Viswanadhan V N 2013 *Current Computer-Aided Drug Design* **9** 482–490 ISSN 1573-4099/1875-6697 URL http://www.eurekaselect.com/node/116577/article

[22] Ma J, Sheridan R P, Liaw A, Dahl G E and Svetnik V 2015 *Journal of Chemical Information and Modeling* **55** 263–274 pMID: 25635324 (*Preprint* http://dx.doi.org/10.1021/ci500747n) URL http://dx.doi.org/10.1021/ci500747n

[23] Kovalenko A and Hirata F 1999 *J. Chem. Phys.* **110** 10095–10112

[24] Luchko T, Gusarov S, Roe D R, Simmerling C, Case D A, Tuszynski J and Kovalenko A 2010 *J. Chem. Theory Comput.* **6** 607–624 ISSN 1549-9618 PMID: 20440377 PMCID: PMC2861832 URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2861832/

[25] Chandler D, Mccoy J D and Singer S J 1986 *J. Chem. Phys.* **85** 5971–5976

[26] Kovalenko A and Hirata F 2000 *J. Chem. Phys* **112** 10391–10402 ISSN 0021-9606, 1089-7690 URL http://scitation.aip.org/content/aip/journal/jcp/112/23/10.1063/1.481676

[27] Kovalenko A Three-dimensional rism theory for molecular liquids and solid-liquid interfaces *Molecular Theory of Solvation* Understanding Chemical Reactivity ed Hirata F (Springer Netherlands) pp 169–275 ISBN 978-1-4020-1562-5 edited by F. Hirata

[28] Truchon J F, Pettitt B M and Labute P 2014 *J. Chem. Theory Comput.* **10** 934–941 ISSN 1549-9618 URL http://dx.doi.org/10.1021/ct4009359

[29] Misin M, Fedorov M V and Palmer D S 2016 *J. Phys. Chem. B* **120** 975–983 ISSN 1520-6106 URL http://dx.doi.org/10.1021/acs.jpcb.5b10809

[30] Palmer D S, Misin M, Fedorov M V and Llinas A 2015 *Mol. Pharmaceutics* **12** 3420–3432 ISSN 1543-8384 URL http://dx.doi.org/10.1021/acs.molpharmaceut.5b00441

[31] Perkyns J S and Pettitt M B 1992 *Chemical Physics Letters* **190** 626–630 ISSN 0009-2614 URL http://www.sciencedirect.com/science/article/pii/000926149285201K

[32] Kovalenko A and Hirata F 2000 *The Journal of Chemical Physics* **113** 2793–2805

[33] Arnot J and Gobas F 2003 *QSAR & Combinatorial Science* **22** 337–345 ISSN 1611-0218 URL http://dx.doi.org/10.1002/qsar.200390023

[34] OECD 2012 Paris URL /content/book/9789264185296-en

[35] Arnot J A and Gobas F A 2006 *Environmental Reviews* **14** 257–297

[36] Papa E, Dearden J and Gramatica P 2007 *Chemosphere* **67** 351–358 ISSN 0045-6535 URL http://www.sciencedirect.com/science/article/pii/S0045653506012732

[37] Gramatica P and Papa E 2005 *QSAR & Combinatorial Science* **24** 953–960 ISSN 1611-0218 URL http://dx.doi.org/10.1002/qsar.200530123

[38] Zhao C, Boriani E, Chana A, Roncaglioni A and Benfenati E 2008 *Chemosphere* **73** 1701–1707 ISSN 0045-6535 URL http://www.sciencedirect.com/science/article/pii/S0045653508011922

[39] 2016 User's guide for t.e.s.t.(toxicity estimation software tool) accessed: May 05, 2017 URL https://www.epa.gov/sites/production/files/2016-05/documents/600r16058.pdf

[40] Dimitrov S, Dimitrova N, Parkerton T, Comber M, Bonnell M and Mekenyan O 2005 *SAR and QSAR in Environmental Research* **16** 531–554 pMID: 16428130 (*Preprint* https://doi.org/10.1080/10659360500474623) URL https://doi.org/10.1080/10659360500474623

[41] Database E B F B G S Accessed: 2017-04-04 URL http://ambit.sourceforge.net/euras/

[42] Rdkit: Open-source cheminformatics accessed: May 05, 2017 URL http://www.rdkit.org

[43] Jean-Paul E, Garrett M and Charlotte D 2012 *Journal of Chemical Information and Modeling* **52** 1146–1158 pMID: 22482737

[44] Rappe A K, Casewit C J, Colwell K S, Goddard W A and Skiff W M 1992 *Journal of the American Chemical Society* **114** 10024–10035 (*Preprint* http://dx.doi.org/10.1021/ja00051a040) URL http://dx.doi.org/10.1021/ja00051a040

[45] Case D, Betz R, Botello-Smith W, Cerutti D, TE Cheatham I, Darden T, Duke R, Giese T, Gohlke H, Goetz A, Homeyer N, Izadi S, Janowski P, Kaus J, Kovalenko A, Lee T, LeGrand S, Li P, Lin C, Luchko T, Luo R, Madej B, Mermelstein D, Merz K, Monard G, Nguyen H, Nguyen H, Omelyan I, Onufriev A, Roe D, Roitberg A, Sagui C, Simmerling C, Swails J, Walker R, Wang J, Wolf R, Wu X, Xiao L, York D and Kollman P 2016 Amber 2016 university of California, San Francisco

[46] Jakalian A, Jack D B and Bayly C I 2002 *Journal of Computational Chemistry* **23** 1623–1641 ISSN 1096-987X URL http://dx.doi.org/10.1002/jcc.10128

[47] The HDF Group 1997-2017 Hierarchical Data Format, version 5 accessed: May 05, 2017 URL http://www.hdfgroup.org/HDF5/

[48] Tokui S, Oono K, Hido S and Clayton J 2015 Chainer: a next-generation open source framework for deep learning *NIPS* URL http://learningsys.org/papers/LearningSys_2015_paper_33.pdf

[49] Golkov V, Skwark M J, Mirchev A, Dikov G, Geanes A R, Mendenhall J, Meiler J and Cremers D 2017 *ArXiv e-prints* (*Preprint* 1704.04039)

[50] He K, Zhang X, Ren S and Sun J 2015 *CoRR* **abs/1502.01852** URL http://arxiv.org/abs/1502.01852

[51] Kingma D P and Ba J 2014 Adam: A method for stochastic optimization accessed: 2017-04-04 (*Preprint* 1412.6980) URL http://arxiv.org/abs/1412.6980v9

[52] Tieleman T and Hinton G 2012 Coursera: Neural networks for machine learning, lecture 6.5 – rmsprop accessed: 2017-04-04 URL http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf

[53] Funk S 2015 Smorms3 - blog entry: Rmsprop loses to smorms3 - beware the epsilon! accessed: 2017-04-04 URL http://sifter.org/simon/journal/20150420.html

[54] Chen T and Guestrin C 2016 *CoRR* **abs/1603.02754** URL http://arxiv.org/abs/1603.02754

[55] Sheridan R P, Wang W M, Liaw A, Ma J and Gifford E M 2016 *Journal of Chemical Information and Modeling* **56** 2353–2360 pMID: 27958738

[56] Kearnes S, McCloskey K, Berndl M, Pande V and Riley P 2016 *Journal of Computer-Aided Molecular Design* **30** 595–608 ISSN 1573-4951 URL http://dx.doi.org/10.1007/s10822-016-9938-8

[57] Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A and Adams R P 2015 *ArXiv e-prints* (*Preprint* 1509.09292)

[58] 2016 Democratizing deep-learning for drug discovery, quantum chemistry, materials science and biology https://github.com/deepchem/deepchem

[59] Sushko I, Novotarskyi S, Körner R, Pandey A K, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko V V, Tanchuk V Y, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin I I, Palyulin V A, Radchenko E V, Welsh W J, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de Sousa J, Zhang Q Y, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V and Tetko I V 2011 *Journal of Computer-Aided Molecular Design* **25** 533–554 ISSN 1573-4951 URL https://doi.org/10.1007/s10822-011-9440-2

[60] Sergiievskyi V, Jeanmairet G, Levesque M and Borgis D 2015 *J. Chem. Phys.* **143** 184116 ISSN 0021-9606, 1089-7690 URL http://scitation.aip.org/content/aip/journal/jcp/143/18/10.1063/1.4935065

[61] Sosnin S 2017 Clever URL dx.doi.org/10.5281/zenodo.835526

[62] Chemaxon 2016 Instant jchem 17.1.16.0 URL http://www.chemaxon.com