SWUS: Active Learning with Structure Weighted Uncertainty Score

Andrea Karlova [1] **Brooks Paige** [1]

Abstract

Active learning has been successfully used in the chemistry to improve the performance of the learner including the out-of-sample generalisation monitoring. The standard query functions utilise the model characteristics such as model uncertainty and related information quantities. While focusing on epistemic uncertainty, the learner utility function often omits the aleatoric uncertainty or exploration of the data manifold structure. In this paper we propose two novel query functions which incorporate the structural information about the chemical diversity of the data. We investigate the performance in comparison to various active learning strategies and under the distributional shifted datasets.

The datasets used in the chemistry contains hidden biases due to various experimental design choices (Thompson et al., 2022). Finding novel lead molecules is subjected to the practical requirements such as cost of experimentally verifying prediction, synthesis-ability or purchase-ability of the molecules rather than optimising the selection for the chemically diverse candidates. While the greedy strategy leads to finding novel binders in short run, it limits the discovery of novel scaffold leads and fails to provide useful learning examples for the predictive model that would provide greater benefits in the long run, (Scalia et al., 2020).

Active learning has been found useful in the chemistry applications (Smith et al., 2018; Yin et al., 2023). It is an adaptive learning paradigm, which iteratively improves the prediction efficiency by testing the new hypotheses about the model performance as a part of the interactive learning process, (MacKay, 1992; Settles, 2009). The learner, represented by the model, queries the unlabeled data with some budget constrain to increase its performance based on getting the new piece of knowledge from the oracle, e.g. lab experiment or queries the foundational model. The instances with higher utility are selected for further labeling allowing the learner to explore the dataset in the biased way to optimise its performance. This, however, does not guarantee that the learner selects the most representative samples or gains the robustness to out-of-sample cases (Nguyen & Smeulders, 2004; Settles, 2009) or domain shifts (Scalia et al., 2020; Yang et al., 2019).

In this paper we investigate the behaviour of the binary classifier on DUD-E and CHEMBL dataset on ligand-decoy classification task and its ability to generalise on the different scaffolds. Our main contribution is in presenting two novel methods for the selection of the structure weighed uncertainty scores: *SWUS: β-uncertainty* and *SWUS: predicted uncertainty*, which utilises the kernel estimation of the molecular or scaffold similarities and weight the uncertainty estimate with the provided score. Our method provides flexible tool which can be used for incorporating the preferences of the wet lab personal or in connection with foundational models (Gupte et al., 2024) and shows robust performance on the domain shift scenarios.

1. Background

1.1. Structural Similarities in Chemical Space

We represent the molecules in the data manifold via its graph-topological representations, circular fingerprints (CFPs) (Rogers & Hahn, 2010), which contain information about molecular features in an accessible way by encoding the topological environment of each atom. Due to direct availability of the sub-structural information, CFPs become common tool in various ligand-based virtual screening (VS) tasks (Cereto-Massagué et al., 2015; Riniker & Landrum, 2013; Hu et al., 2012).

To explore the structure of the chemical space, we use shape description variant of Bemis-Murko (BM) molecular framework (Bemis & Murcko, 1996). The method deconstructs the molecules into sidechains, rings and linkers. We use 'graph framework' where sidechains are removed, all bond types converted to a single bond and all atom types to carbon. This 'generic framework' method extracts cyclical skeletons. The advantage of this extraction is that we characterise the molecules based on the most bare graphical structure.

To assess the similarities between the molecules, we encode

[1]University College London. Correspondence to: Andrea Karlova <a.karlova@ucl.ac.uk>.

Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery, Vienna, Austria.

the molecules, cyclical skeletons, respectively, into the CFPs and use Tanimoto similarity (TS) (Holliday, 2002) to score the graph-topological differences between the molecules. To model the similarity of the data manifold, we use the kernel density (Parzen, 1962) with Tanimoto kernel. This introduces the structure into a prior over the molecular data manifold.

For a virtual library $\mathcal{D}$ with molecules $m_i, i = 1, \ldots, n$, we denote Tanimoto similarity kernel $k(m_i, m_j)$ between two molecules m_i and m_j. Each molecule m_i is scored with the mean with respect to all other molecules in the data manifold as follows:

$$\mathbb{E}p_{mol}(m_i) = \frac{1}{|\mathcal{D}|} \sum_{m_j \in \mathcal{D}} k(m_i, m_j). \tag{1}$$

We apply the same methodology for the cyclical skeletons. The molecules from the library $\mathcal{D}$ are converted to the library of cyclical skeletons $\mathcal{D}_{csk}$ with skeletons $s_i, i = 1, \ldots, n$. Tanimoto kernel is placed over the manifold of the cyclical skeletons:

$$\mathbb{E}p_{csk}(s_i) = \frac{1}{|\mathcal{D}_{csk}|} \sum_{s_j \in \mathcal{D}_{csk}} k(s_i, s_j). \tag{2}$$

1.2. Active Learning with Manifold Exploration

The learner selects the points that have the highest utility to it for exploiting the performance instead of weighting in the exploration. It improves its performance based on its internal model preferences, e.g. those which are the closest to the classification boundary, rather than taking into consideration the underlying structure of the data (Nguyen & Smeulders, 2004). Incorporating the structural information into the query is beneficial from both, the real-life application and learner's performance perspective. It allows to better control whether the data from the similar or different underlying representations are selected. Modifying the query function such that the learner has an option to explore the underlying data structure gives potential benefits, such as mechanism for encoding preferences for diverse scaffolds, chemical groups or pharmacophore-like features.

2. Query with the Chemical Space Structural Awareness

Our model is a binary classifier which assigns labels $\{0, 1\}$ distinguish between the ligands and decoys. To model the uncertainty, we use probabilistic model with the posterior distribution $p(y, \theta|x) = p(y|\theta, x)p(\theta)$ modelling the epistemic uncertainty, where $p(\theta)$ is the prior over the model parameters $\theta \in \Theta$ and the marginal prediction of the model is computed as $p(y|x) = \mathbb{E}_{p(\theta)}p(y|x, \theta)$. (Cohn et al., 1994; Nguyen & Smeulders, 2004) suggests to select samples that minise the future classification error. Using probabilistic model we can quantify the outcome as: $\int_{\mathcal{D}} \mathbb{E}_{p(y|x)}|\hat{y}(x) - y(x)|^2 p(x)\mathrm{d}x$, where $\hat{y}(x)$ is the model prediction and $y(x)$ is the true label from the oracle. We quantify the uncertainty using the least confident score: $\mathcal{U}(x, \theta) = 1 - |p(\hat{y}(x) = 1|\theta, x) - p(\hat{y}(x) = 0|\theta, x)|$. Note that this score can be replaced with information gain or other utility function that quantifies the model performance based on the state of its parameters θ subjected to given data.

Next we introduce the Structure-weighted uncertainty score:

$$SWUS_\theta(m_i) = \mathcal{U}(m_i, \theta) \left[\frac{1}{|\mathcal{D}|} \sum_{m_j \in \mathcal{D}} k(m_i, m_j) \right]^\beta, \tag{3}$$

where m_i is a molecule and $k(\cdot, \cdot)$ is TS kernel applied on molecules. Note m_i can be replaced with scaffolds s_i, depending on the task requirements. Parameter β acts as weight preference of the information from manifold structure, (Settles, 2009).

The initial experiments suggest rather erratic behaviour of the learners when the data pool is queried one-by-one or in a small batches. This is non-desirable. In fact, we can expect the learner to first aim to improve its own performance and when it starts to stagnate, it seeks the new source of information (Nguyen & Smeulders, 2004). We propose a method, where the learner decides about the amount of the information from the manifold structure which it wants to utilise. For the grid of values $\beta_l, l = 1, \ldots, K$, the learner scores the molecules using (3), narrows down K candidate molecules and selects molecule with corresponding weight preference β_l which indicates the highest uncertainty about the prediction.

We introduce two algorithms according which the learner decides about the weight β_l of the structural information. The first method, described in Algorithm 1, simply selects the β_l with the highest $SWUS_\theta(\beta_l)$ score. The advantage is a computational speed while giving learner the preference option based on the similarity information.

The second method is computationally more expensive and is based on the premise, that the learner locally carries small scale experiment based on which it decides whether to include the candidate point. The learner proceeds like in Algorithm 1 and selects K candidate molecules. Next, the learner constructs K scenarios where it explores situation in which the oracle returns label 0 and in which oracle returns label 1. For these selected K points, the learner constructs $2K$ candidate training sets with labels 0 and 1, re-train itself to these scenario datasets and assess its confidence into assigned scores. This allows the learner to explore in more robust way the possible future scenario by computing the expected future uncertainty. See Algorithm 2 for the detailed description.

Algorithm 1 SWUS: $\beta-$ uncertainty

Initialize base learner (training set $\mathcal{M}^0$, classifier θ^0)
for $iteration\ i = 1, 2, \ldots$ **do**
  for $\beta_l - learner:\ l = 1, 2, \ldots, K$ **do**
    Compute score $SWUS_{\theta^{i-1}}(\beta_l)$
  end for
  Choose $m^i = \text{argmax}\ SWUS_{\theta^{i-1}}(\beta_l)$
  Query label for m^i
  Add m^i to the training set: $\mathcal{M}^i = \mathcal{M}^{i-1} \cup \{m^i\}$
  Train learner (get θ^i)
end for

Algorithm 2 SWUS: predicted uncertainty

Initialize base learner (training set $\mathcal{M}^0$, classifier θ^0)
for $iteration\ i = 1, 2, \ldots$ **do**
  for $\beta_l - learner:\ l = 1, 2, \ldots, K$ **do**
    Compute score $SWUS_{\theta^{i-1}}(\beta_l)$
    Choose $m_l^i = \text{argmax}\ SWUS_{\theta^{i-1}}(\beta_l)$
    Classify m_l^i to get the decoy prob. $p_{\theta^{i-1}}$
    Add m_l^i to $\beta_l - learner$'s training set:
      $\mathcal{M}_l^i = \mathcal{M}^{i-1} \cup \{m_l^i\}$
    Train binary classifier α assuming m_l^i has label:
      $y_l^i = 0$ (decoy).
    Compute the new decoy prob., p_α
    Train binary classifier γ assuming m_l^i has label:
      $y_l^i = 1$ (ligand).
    Compute the alternative new decoy prob., p_γ
    Compute the predicted uncertainty:
    $U_l = 1 - |\{p_{\theta^{i-1}}|2p_\alpha - 1| - (1 - p_{\theta^{i-1}})|2p_\gamma - 1|\}|$
  end for
  Select $l^* = \text{argmax}\ U_l$
  Add $m_{l^*}^i$ to the training set: $\mathcal{M}^i = \mathcal{M}^{i-1} \cup \{m_{l^*}^i\}$
  Train classifier (get θ^i)
end for

3. Experiments

3.1. Datasets

To demonstrate the performance of our method, we selected adenosine A2a G protein-coupled receptor (AA2AR) from two closely related datasets: DUD-E and CHEMBL. CHEMBL (Mendez et al., 2018) is publicly available manually curated database of bioactive molecules. DUD-E (Mysinger et al., 2012), the Directory of Useful Decoys - Enhanced, originally designed for assessing the docking performance.

The ligands associated to each target were selected from CHEMBL09 (Mendez et al., 2018). The criteria used for the selection guarantee reasonable morphological properties and potential bioactivity, e.g. molecules included if the binding affinities are better than 1μM, molecular weight is less than 600, the limit of 20 on the number of the rotatable bonds. To increase the scaffold diversity, the selected ligands were clustered using BM atomic framework. From these scaffold clusters, only representative ligands, e.g. with highest binding affinity are included. For each clustered ligand, there is matched 50 decoys from ZINC database (Irwin et al., 2012). The decoys are either experimentally known or generated to match the chemical properties of the ligands, such as molecular weight, logP, number of rotatable bonds, hydrogen donors and acceptors. The topological properties of the decoys are however different from the ligands. Therefore, by design, the clustered ligands together with matched decoys and associated to various target types are topologically and chemically diverse and curated for model training. We use this dataset for sampling the balanced training set for the learner and to construct the highly dis-balanced test-set.

To compare the performance of the selected approaches, we use two types of test sets, both originating from CHEMBL. The first one contains all ligands which are included in DUD-E ligands set but were discarded during BM clustering. We can expect that the scaffolds included in this set are somewhat similar to the scaffolds in the training set, while the binding activities may be potentially weaker. As ligands in DUD-E dataset are from CHEMBL09, we select additional ligands from the CHEMBL25. This mimics the time-split and includes novel scaffolds. Our selection of the test sets allow to quantify the performance of the active learners under two different criteria: 1) assesses the ability of the method to identify the bioactive molecules which are structurally somewhat similar to those included during the training, 2) allows to measure the performance of the active learner based on the diversity of the structural properties, 3) tests the performance of the classifier on the disbalanced dataset indicating distributional shift and realistic VS task. All overlapping molecules between the datasets were excluded.

	activity	cyclical skeleton	smile
DUD-E	ligand	238	482
	decoy	7166	31547
CHEMBL 09	ligand	438	2614
CHEMBL 25	ligand	540	1894

Table 1: Adenosine A2a receptor: the profile of the library available from CHEMBL 09, CHEMBL 25 and DUD-E. While there is a reasonable diversity in the reported ligands, the generated decoys contains many similar cyclical skeletons.

3.2. Bioactivity Classification

We use Random Forest as a bio-activity classifier, as it is fast to train robust benchmark, which is often used in practise.

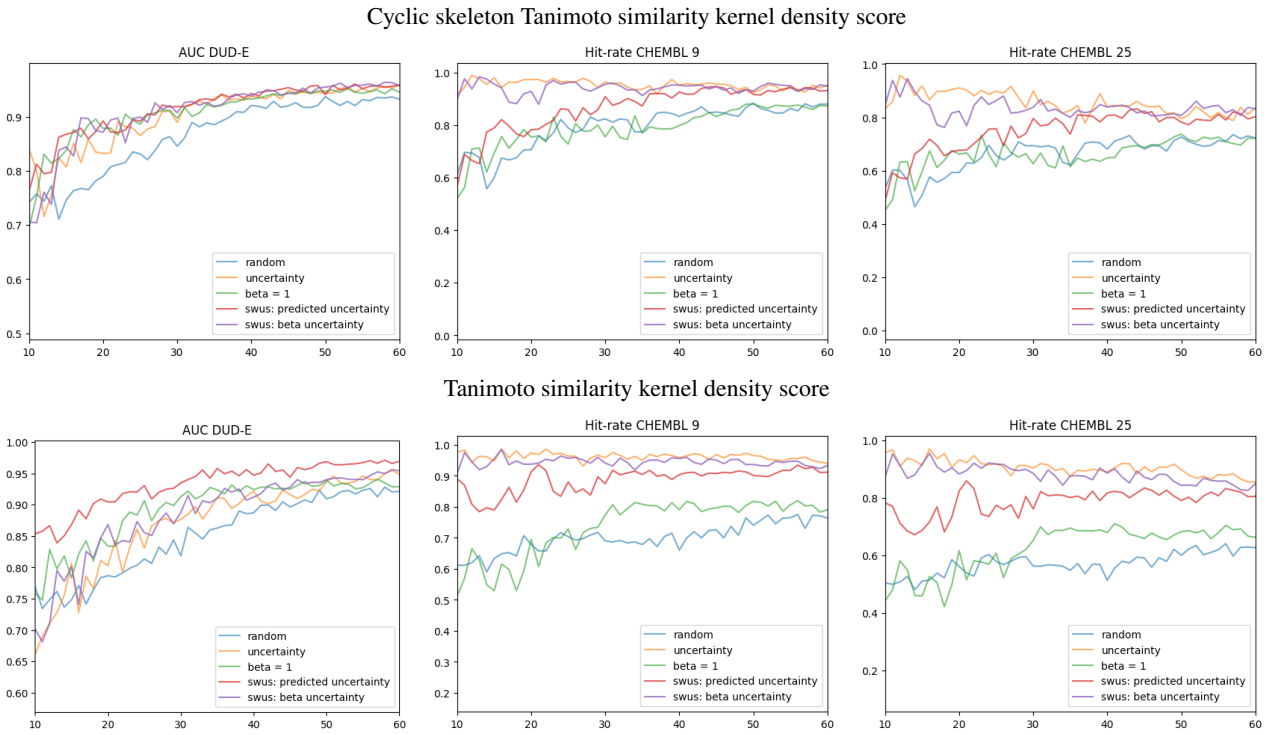

Figure 1: **Bio-activity Classifier for AA2AR receptor.** We provide comparison of 5 different query functions for the RF learner on 3 different test-sets. The base learner was trained on the balanced set of 5 ligands and 5 decoys. After each query, we evaluate the learner on three different test sets: the first test-set is the dis-balanced DUD-E sample of 95 ligands and 6312 decoys; the second test contains ligands from CHEMBL9, the third test set contains ligands from CHEMBL 25. The plotted line for each query is the mean of 5 repetitions. The learner has a budget of 50 queries. We provide comparison to two different structure weights: one is based on the cyclic skeleton TS kernel density estimate, the second is based on the TS kernel density estimate without any initial clustering.

We use RDKit's implementation of ECFP with radius 3 and length 1024 as features for the learner.

We randomly select 20% of the DUD-E dataset as a test set. As DUD-E is highly dis-balanced, the test set contains around 1.5% of ligands only. We initiate the training set by sampling 5 actives and 5 decoys from the remaining 80% held-out. 80% held-out represents the library on which we perform VS. We compute the TS kernel density on the molecules, the cyclical skeletons, respectively, for each molecule of the library, see Table 1 for the profile of the datasets.

We use pool-based sampling scenario and use random sampling as a benchmark. We incorporate select the grid of β values: $[-1, -0.5, 0, 0.5, 1, 1.5, 2]$, construct 'SWUS - predicted uncertainty' query and 'SWUS - beta uncertainty'. We compare the query performance with the uncertainty score without any structural information and the uncertainty score weighted by the structural score ($\beta = 1$). While the structural score with $\beta = 1$ does not provide visible benefit, the 'SWUS -predicted uncertainty' demonstrates robust performance with respect to distributional shift within the given budget. Its performance does not deteriorate on the distributional-shifted test sets while including new data point from the unlabeled pool. While transforming the molecules to cyclic skeletons doesn't provide computational advantage to computing TS for original molecules, there is a clear computational advantage when computing similarity weights. See Figure 1 for the results.

4. Discussion

We present two query strategies that allow to perform the VS of the molecular libraries to identify the bio-active candidates based on the molecular similarity and uncertainty of the learner. We show how the structural preferences can be encoded into the query function while giving the learner the flexibility to decide how useful is the structural information for its performance. Our method demonstrates better robustness under the distributional shift, which is feasible for wet-lab experiments or in a connection with the foundational models.

References

Bemis, G. W. and Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, January 1996. ISSN 1520-4804. doi: 10.1021/jm9602928. URL http://dx.doi.org/10.1021/jm9602928.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, jan 2015. doi: 10.1016/j.ymeth.2014.08.005.

Cohn, D., Ghahramani, Z., and Jordan, M. Active learning with statistical models. In Tesauro, G., Touretzky, D., and Leen, T. (eds.), *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/file/7f975a56c761db6506eca0b37ce6ec87-Paper.pdf.

Gupte, S. R., Aklilu, J., Nirschl, J., and Yeung-Levy, S. Revisiting active learning in the era of vision foundation models. *ArXiv*, abs/2401.14555, 2024. URL https://api.semanticscholar.org/CorpusID:267301205.

Holliday. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2d fragment bit-strings. *Combinatorial Chemistry amp; High Throughput Screening*, 5(2), 2002. ISSN 1386-2073. doi: 10.2174/1386207024607338. URL http://dx.doi.org/10.2174/1386207024607338.

Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G., and Tang, Y. Performance evaluation of 2d fingerprint and 3d shape similarity methods in virtual screening. *Journal of Chemical Information and Modeling*, 52(5):1103–1113, may 2012. doi: 10.1021/ci300030u.

Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

MacKay, D. J. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., Veij, M. D., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., Hersey, A., and Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47 (D1):D930–D940, nov 2018. doi: 10.1093/nar/gky1075.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012. doi: 10.1021/jm300687e. URL https://doi.org/10.1021/jm300687e.

Nguyen, H. T. and Smeulders, A. W. M. Active learning using pre-clustering. In *International Conference on Machine Learning*, pp. 623–630, 2004. URL https://ivi.fnwi.uva.nl/isis/publications/2004/NguyenICML2004.

Parzen, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33 (3):1065–1076, September 1962. ISSN 0003-4851. doi: 10.1214/aoms/1177704472. URL http://dx.doi.org/10.1214/aoms/1177704472.

Riniker, S. and Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1), may 2013. doi: 10.1186/1758-2946-5-26.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, apr 2010. doi: 10.1021/ci100050t.

Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P., and Green, W. H. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 60(6):2697–2717, 2020. doi: 10.1021/acs.jcim.9b00975. URL https://doi.org/10.1021/acs.jcim.9b00975.

Settles, B. Active learning literature survey. 2009.

Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24), 2018. ISSN 1089-7690. doi: 10.1063/1.5023802. URL http://dx.doi.org/10.1063/1.5023802.

Thompson, J., Walters, W. P., Feng, J. A., Pabon, N. A., Xu, H., Maser, M., Goldman, B. B., Moustakas, D., Schmidt, M., and York, F. Optimizing active learning for free energy calculations. *Artificial Intelligence in the Life Sciences*, 2:100050, 2022. ISSN 2667-3185. doi: https://doi.org/10.1016/j.ailsci.2022.100050. URL https://www.sciencedirect.com/science/article/pii/S2667318522000204.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and

Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. ISSN 1549-960X. doi: 10.1021/acs.jcim.9b00237. URL http://dx.doi.org/10.1021/acs.jcim.9b00237.

Yin, T., Panapitiya, G., Coda, E. D., and Saldanha, E. G. Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. *Journal of Cheminformatics*, 15(1), 2023. ISSN 1758-2946. doi: 10.1186/s13321-023-00753-5. URL http://dx.doi.org/10.1186/s13321-023-00753-5.