Leveraging binding-site structure for drug discovery with point-cloud methods

Vincent Mallet
Department of Computer Science
McGill University
vincent.mallet@mail.mcgill.ca

Carlos G. Oliver
Department of Computer Science
McGill University
carlos.gonzalezoliver@mail.mcgill.ca

Nicolas Moitessier
Department of Chemistry
McGill University
nicolas.moitessier@mcgill.ca

Jérôme Waldispühl
Department of Computer Science
McGill University
jerome.waldispuhl@mcgill.ca

Abstract

Computational drug discovery strategies can be broadly placed in two categories: ligand-based methods which identify novel molecules by similarity with known ligands, and structure-based methods which predict molecules with high-affinity to a given 3D structure (e.g. a protein). However, ligand-based methods do not leverage information about the binding site, and structure-based approaches rely on the knowledge of a finite set of ligands binding the target. In this work, we introduce `TarLig`, a novel approach that aims to bridge the gap between ligand and structure-based approaches. We use the 3D structure of the binding site as input to a model which predicts the ligand preferences of the binding site. The resulting predictions could then offer promising seeds and constraints in the chemical space search, based on the binding site structure. `TarLig` outperforms standard models by introducing a data-alignment and augmentation technique. The recent popularity of Volumetric 3DCNN pipelines in structural bioinformatics suggests that this extra step could help a wide range of methods to improve their results with minimal modifications.

1 Introduction

***In silico* drug discovery** The majority of drugs are small organic molecules that alter cellular mechanisms upon binding to key bio-molecules such as proteins. Identifying molecules with a desired set of properties is known as the drug discovery process. This process often relies on an iterative, tedious and expensive process. Because most molecules are inactive or lack specificity toward a given function and binding site, Virtual Screening (VS) tools have emerged as an essential tool to identify the most promising candidates molecules prior to synthesis and/or experimental evaluation. Given a initial set of generic small molecules (i.e. the library), VS methods output a subset of this library enriched with active compounds. Another use of these computational methods is to predict off-target binding (binding to a non-desired site) to assess the binding specificity to the molecular target. This information is in particular helpful to anticipate potentially toxic compounds.

Structure-based approaches Most VS methods broadly fall under two categories [1]. Structure-based approaches use the 3D structure of the protein binding site to identify promising candidate ligands (e.g., enzyme inhibitors, receptor antagonists,...). If the binding site is not known, several tools are able to detect the binding sites of a protein [2, 3]. Knowledge of the binding site shape and

requirements from ligand-protein co-crystals can be exploited to dock ligands into this binding site. This process is computationally demanding (i.e., from a few seconds to a few minutes per ligand) and consists in two phases: one that explores possible binding orientations of the ligands in the binding site and another that ranks them by affinity. This approach has been the subject of detailed studies [4, 5]. Some machine learning approaches can be used to do the ranking step or even bypass the docking step and directly compute an affinity score[6]. However, increased accuracy comes at the cost of trying a larger number of orientations and therefore cannot be easily accelerated. The shortcoming of such methods is that they take as input both a ligand and the target protein and thus they are restrained to a limited set of ligands. This leads to limitations because this finite set has already been extensively studied and patented. Additionally, these methods are often computationally expensive.

Ligand-based approaches   Ligand-based approaches assume that ligands structurally similar to active ligands will also be active[7, 8]. While this concept ignores the actual biological binding site explicitly, it implicitly considers the potential interactions and takes advantage of experimental binding affinities of reference ligands to search for new ones. Interest towards generative models in this ligand-centered task is rising[9]. A major asset of this class of methods is the ability to identify potential ligands that lie outside of the patented molecular space. These models often rely on unsupervised methods that use a similar strategies to `word2vec` in the molecular space [10]. The latent space induced by such methods is conveniently prone to algebra. Notably, we can use optimization processes in the latent space to fine-tune a seed for ligand druggability [11, 12]. These methods are becoming increasingly popular but struggle to use binding site information, relying instead on extensive experimental assays to reference ligands. Although it is not a fundamental bottleneck for target prediction tasks, this is a major limitation for off-target predictions.

Our contribution   In this paper, we propose to integrate binding site information into a ligand-based approach. For this purpose, we introduce a new task: we use protein structure information to quickly identify regions of interest in ligand space and use this knowledge as a potential guide in the ligand generative process. More specifically, we use the 3D structure of a protein binding site as input and predict a latent space embedding of a ligand. To our knowledge, this kind of approach was only conducted independently in a preprint [13] that used graphs to represent the pocket. Our model learns a rotation invariant representation for the 3D input using an alignment and augmentation step, showing superior results against 3D steerable and classic Volumetric CNNs pipelines. We believe this step is relevant for 3D shape comparison in general and could help other Volumetric CNN models improve their results.

2   Related methods and data representation

2.1   Ligand representation

Sparse representations   Ligands can be represented as graphs, strings, sets of subgroups or 3D structures. All of these representations can in turn be embedded into vectors. Molecular fingerprints are a commonly used representation defined as a bit-string encoding the presence or absence of certain relevant chemical fragments according to domain experts [14, 15]. Such representations have the usual advantages and drawbacks of hand-featured representations: they do not rely on abundant data and are easily interpretable. However, they are also not compact and potentially miss important features of the chemical space. The largest chemical compounds database (REAL database) contains 10^9 compounds that could be represented using binary vectors of dimension $log_2(10^9) \sim 30$ instead of the fingerprints of length 1024 used by the ECPF6 [15] representations. Thus, hand-crafted, binary embeddings and the bit-to-bit distance (Tanimoto coefficient) span a sparse discontinuous space with regards to the space usually explored by chemists. Continuous embeddings allow a more compact embedding and better preservation of chemical similarity that are the basis of all potential machine learning algorithms.

Data-driven representations   Data-driven representations arose in 2015 and were first applied to graph representations of ligands [16, 17]. Next, unsupervised methods of auto encoding were applied to larger data sets using the `SMILES` representation of millions of compounds available in large databases such as ZINC[18] with a paper leveraging similar ideas as `word2vec` to get embeddings

of molecules [10]. The authors were the first to introduce the idea that the latent space could serve as a good molecular representation for other learning tasks such as bio-activity prediction. A lot of work was conducted in this domain to adapt it to the use of VAEs or GANs[12] and we chose to base our work on a transitional auto encoder [19] trained from `SMILES` to canonical `SMILES` on over one billion data points. We will represent our ligands with the latent space representation of this model.

2.2 Protein representations

The functional level of proteins and of most biological objects resides in their 3D structure and is composed of finite number of elements. We preferred the atomic resolution over the amino-acid resolution as the latter cannot properly model the orientation of each amino acid and side-chain details, two major factors in protein-ligand binding. Therefore the binding sites are fundamentally modeled as a point cloud of atoms. Several deterministic embedding tools exist that turn these objects into vectors [20] but for reasons similar to ligands, learned representations are more promising [21].

2.3 Point cloud networks

While we chose to model protein binding sites as a point cloud - ie a matrix of coordinates associated with atom types - for biological motivations, there is no well established way to process such data. The challenges are choosing a translation and rotation invariant coordinate system to express our data in and an order to consider our points. The point ordering is partially solved by a framework such as PointNet [22] using MaxPooling over all points. The most common solution to the ordering challenge is to consider this cloud as a sparse 3D image and to leverage the CNN frameworks, which consists in the Volumetric CNN framework. Treating objects as images fails to respect their rotational invariance and leads to very sparse inputs. For 2D images there is usually an orientation convention but there is no such native pose of a 3D object. This makes data augmentation much more difficult. Moreover as argued in [23] the data augmentation is much more challenging in 3D than in 2D.

For these reasons, recent works try to extend the invariance properties of CNNs. In 2D, some work has shown superior performance including rotational and symmetry invariances [24, 25]. In 3D, recent models manage to include rotational invariance[26] or equivariance[23, 27]. These models have the ability to leverage the data properties. However they do not usually have clear convergence properties. Also since they have not been extensively used and studied, they are harder to use and to train than usual networks. We implemented a Volumetric pipeline with modifications and we chose to also use steerable 3DCNN[27] because they required only minimal modification from the Volumetric pipeline

3 Methods

3.1 Data and performance metrics

The protein binding sites we used were extracted from the PDB[28] from bound examples. The extraction process consisted in taking bound protein structures, extracting at all amino-acids around the ligand and extending the radius around each of its atom until a certain radius or a certain number of neighbors is reached . We removed the metabolites in order to focus on larger, more interesting binding molecules. We then removed the binding sites that we considered duplicates using sequence identity cutoff and removing several examples of the same ligand bound to the same protein.

Table 1: Data extracted from the PDB

Number of proteins	Size in Angstrom	Number of ligands
30964	40 * 30 * 30	3362

As a metric we will often use the Mean Square Error (MSE) and the Enrichment factor (EF). EF at the ith level consist in the rate of active ligands in a subset of size $i\%$ of the original data set, normalized by the rate of actives in the whole set. EF was designed to mimic the true use case of virtual screening where chemists only test a small subset and hope to have as many actives as possible.

We used the DUDE [29] database that is the standard for assessing enrichment factors. The DUDE database consists in 102 targets and their associated set of actives. They then create 50 decoys per actives, preserving the physico-chemical features, but making the structure different, to compromise the binding.

3.2 Principal axis alignment

We also offer an alternate strategy to bypass the problem of rotation invariance. For 3D objects embedded in images, we can compute the PCA as a way to find a consensus way to align these objects addressing the lack of native pose problem. Computing these axis is costly but it can be done as a preprocessing step and enables us to bypass the registration problem. Since the eigenvectors have no canonical sign, there remains an ambiguity regarding the pose but for a k-dimensional space, it is reduced to $2^k = 8$ possible poses for a 3D object.

We first validated this idea to use the principal axis of 3D objects by applying it to a small molecules comparison method : USRCAT [30]. This method uses the successive first moments of the distribution of distances to references points as an embedding for a given molecule. As an example, the first feature of a set of points would be its mean distance to its centroid. We used the PCA-based approach to get orientation independent reference points (at a given distance in angstrom from the centroid, following the eigenvectors axis). We used the benchmark used in the latest version of this method and show superior performance in term of enrichment factor.

We then used this idea to represent our binding sites all aligned on the same grid. A side advantage of this alignment step is that it enables us to fit all object on a smaller grid since the dimensions with the highest variance usually correspond to the ones with largest values - with the exception of outliers. Using this framework solves the translation problem and enables us to drastically reduce the problem of representing rotations of point clouds. We can use data augmentation to put all 8 flipped views in the data set. We can then think our network might learn the invariance to flips. If we use a permutation invariant reduction operation such as the average on the flips, we make the network invariant to rotations.

To explicitly enforce invariance to rotations for each prediction, we need to tie weights that are symmetric with respect to one of the axes of the 3D grid. We can also present batches containing the 8 poses and the batch-averaged gradients will then be symmetric. This is an engineering fix that makes the implementation straightforward. Further implementation of this model should include the weight tying. This would reduce both the batch size and the number of parameters. However, since the number of possible models is the same, we should be getting similar results with such constraints actively enforced.

This simple alignment could be an easy yet efficient engineering solution to help building rotation invariant networks. However this approach has a few caveats : we need to compute the PCA of the point cloud for every evaluation of the networks, which could be challenging for high number of points - but not for molecular applications.

We used the pipeline presented in **Figure 1**.

4 Results and discussion

4.1 Alignment Step Validation

To assess the impact of the alignment step with a side application, we competed against USRCAT[30] using reference points derived from the PCA. We computed EF at different thresholds for each of the DUDE[29] binding sites and averaged the results. We obtain similar yet significantly superior results ($\sim 9\%$). The limited improvement can be explained by the method reaching its limits, as explained in [30] but is enough to show that using points derived from the PCA is a promising solution for representing 3D objects.
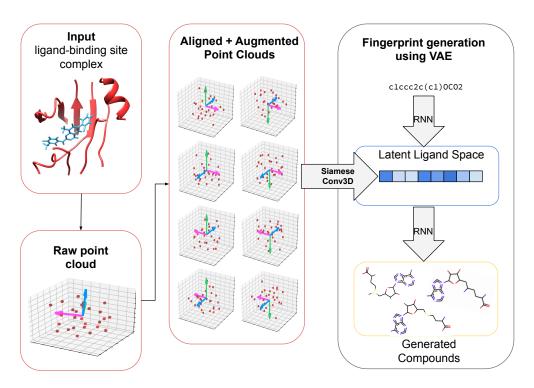
Figure 1: Pipeline used : a binding site (in red) is extracted around a ligand (blue) and turned into a point cloud. This point cloud is aligned to its principal axis and turned into 8 volumetric images. These representations are then used to learn the latent code of the extracted ligand

Table 2: EF computed with different threshold averaged over the DUDE database (the ratio vs our method)

Method	$EF_{0.01}$	$EF_{0.0025}$
UFSR	7.97 (0.45)	4.50 (0.46)
USRCAT	16.50 (0.91)	9.14 (0.94)
USRPCAT	17.51 (1)	9.97 (1)

4.2 3DCNN Results - Alignment step

We now come back to the task of predicting latent representation of ligands from the structure of a binding site. We have trained a number of models to assess the impact of :

- Aligning the data to the PCA axes

- Augmenting the data set with flips

- Using batching or siamese model to enforce invariance

- Different CNN settings with more or less parameters

See **Section 5** for more details on the implementation and infrastructure architecture and parameters.

As a control experiment, we trained our model on shuffled labels. We benchmarked against steerable CNN implementations[27]. We also wanted to assess the bagging effect of models, so we show the average error on each view of a 3D image in addition to the bagged error.

5

Table 3: Mean square error (MSE) between true and predicted ligands.

Method	Best test MSE	Bagged test MSE	Final train MSE	Time to train
`TarLig` shuffled	0.210	0.206	0.07	6h
`TarLig`	0.108	0.108	0.038	1h
`TarLig` flips	0.095	0.092	0.04	6h
`TarLig` PCA	0.103	0.103	0.036	**1h**
`TarLig` PCA flips	0.089	**0.085**	0.033	5h
`TarLig` batched flips	0.095	0.088	0.055	5h
`TarLig` siamese	**0.088**	0.088	**0.023**	4h
Small `TarLig` flips	0.099	0.089	0.035	5h
Se3cnn flips	0.18	0.16	0.091	35h
Small se3cnn flips	Diverged	N/A	0.17	19h

We see in **Table 3** that our model is able to learn, and performs significantly better than the shuffled control. We have the expected result with regards to the PCA alignment : aligning all binding sites results in a higher score. Moreover, these results were obtained using early stopping, but considering the test metric curves, we see that the learning is a lot less stable with unaligned data, i.e. the model is harder to train and the best test error is actually over-fitting on the test set (data not shown). Therefore, we conclude that our alignment strategy is beneficial. The data augmentation also has the expected effect of helping the learning.
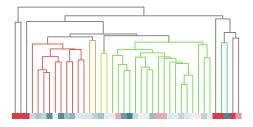
The best model is the one that does not enforce the rotational invariance. Putting all flips in the same batch results in the same value of convergence but takes longer to reach it. This can be interpreted as an 'effective batch size' effect : we enforce rotational invariance at the cost of showing eight times less binding sites in each batch batch (for a constant batch size). The siamese model only pushes the bagged average to the right vector while the other push each individual prediction to the label.
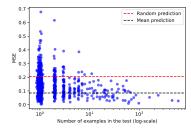
All architectures we trained behaved very similarly. We report the `TarLig` one that achieved the best, yet comparable, results to the ones we trained after, including smaller versions. The trends we just described concerning the impact of aligning the pocket, augmenting the data or using it in a siamese architecture were found to be true across all the models we trained.

The se3cnn network showed poor performance overall and we could not manage to bring its performance close to our results. We think it is a big limitation of such networks, despite having appealing theoretical properties, they are much harder to train and use which is a barrier to fully leveraging their power. The alignment step enabled learning for this task and we think that combined with the volumetric approach and some of the refinements developed for this approach, it could be an efficient and much easier way to use 3D data.

4.3 Performance by ligand family

Next, we investigate the behaviour of our model beyond a simple average MSE score. To do so we broke down our results per ligand and clustered the ligands into a dendrogram to measure the effect of the distribution of ligands in the dataset on performance (Figure 2).

(a) Performance by ligand, with clustering of the ligands

(b) Performance by ligand, split by the number of examples in the set

Figure 2: Performance of the model detailed in the output space

Overall the model performs well across ligand families. The points on the extreme left and right of the dendrogram are ligands that lie far from the main clusters are thus sparsely populated. **Fig. 2** illustrates, naturally, that performance improves as more binding sites for a ligand are obtained. Regardless, we still have a better prediction than random for most points, including those with few or even one example.
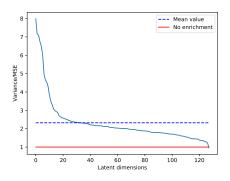
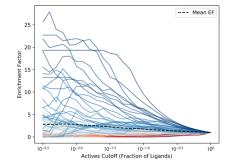### 4.4 Ligand space reduction to enhance structure-based sampling

We now use the average MSE for each output dimension to determine which dimensions are better predicted and as a rough estimate of the uncertainty bounds of our method. One can compare the MSE per-dimension to the variance of the data (the MSE always returning the average). If the MSE is small compared to the variance, we can say that the model has learned to predict this dimension reasonably well (Figure 3a).

For most dimensions, the reduction is near two-fold over random, yet roughly 30 dimensions experience a reduction of up to 8-fold. To contextualize the usefulness of this reduction, we can think of structure-based approaches as taking samples in the chemical space and docking them against the binding site by assessing their affinity. The variance of this sampling is reflected by the variety of compounds observed in the PDB. The MSE defines a high dimensional box in the latent space, where the affinity is high. If we consider uniform sampling the latent space, we would need an expected 10^{42} samples (computed value by the product of bins possibility in each dimension) to have a point fall in this box. while the usual number of compounds that can be handled by structure-based methods is usually only 10^8 molecules. Due to the curse of dimensionality, even though the result is only giving on average a 2-fold enhancement in each dimension, our method is theoretically able to suggest a point unreachable with the usual structure-based approaches.

4.5 Case study and enrichment factor

In this part, we use DUDE [29] as an external database and evaluate the performance of our techniques w.r.t. the enrichment factor. Although we do not aim to match the performance of well-established methods with this settings, it offers an interesting perspective of the potential of our approach. Ligand-based methods look for similarity between actives compounds that exist in the data set, while structure-based methods explicitly evaluate all candidates. By contrast, we do only one prediction. A more complete validation of the quality of our predictions would be to probe their neighborhood. Here, we are conducting two experiments. First, we want to know if we can correctly identify the sub-region of space with given properties. Next, we aim to determine if this sub-region is closer to active compounds than decoys. We make our prediction using the 3D structure of the binding site and compute the distance to our prediction. To get an EF score, we then sort actives and decoys compounds by distance to this prediction.

(a) Variance reduction in the latent space per latent dimension

(b) Enrichment factor at different thresholds for each target of the DUDE Database

Our first results shows that the average distance from a prediction to its associated ligand has an 'MSE' value of 0.13, which is lower than the random one of 0.2. We conclude that our methods are able to identify regions of interest for this task. Importantly, we get an enrichment for most targets,

which means that our model is able to make a prediction closer to the region of actives compounds. However, it seems that inactive compounds do not lie in a well separated space. Thus, the closest points of the predictions necessarily include decoys. A strategy to address this caveat would be developing other embeddings that separates compounds based on affinity.

5 Conclusion and future work

In this paper, we show that learning a representation of known binding sites enables us to predict sub-regions of the ligand space with improved binding potential. These results support more extensive use of binding site information within ligand generative models and suggest novel avenues for improvement for the computational drug discovery pipelines. Additional binding information such as experimental affinity could further enhance results and help overcome bias of learning only on co-crystallized complexes.

The alignment step within binding sites that helps us to amplify the signal, is a simple yet promising solution to represent 3D structures. Eventually, the binding site representations could be improved through a self-supervised pipeline leveraging the increasing amount of structural data available for proteins.

Finally, several lines of work are already developed to improve generative models for ligands. In this context, our method offers promising perspectives to generate relevant seed and help for further optimization [31].

Code and Setup

We used a standard model with 7 convolution layers and 2 fully connected. The detailed architecture can be found on GitHub and is described in the `models/SmallC3D.py` class. One can also run `python main.py -summary` to get a summary of the model architecture and parameters that were used. Training was performed on 4 NVIDIA P100 Pascal GPU using 20 Broadwell cores.

References

[1] Paul C. D. Hawkins, A. Geoffrey Skillman, and Anthony Nicholls. Comparison of shape-matching and docking as virtual screening tools. Journal of Medicinal Chemistry, 50(1):74–82, 2007. PMID: 17201411.

[2] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics, 10(1):168, 2009.

[3] J. Jiménez, S. Doerr, G. Martínez-Rosell, A. S. Rose, and G. De Fabritiis. DeepSite: protein-binding site predictor using 3d-convolutional neural networks. Bioinformatics, 33(19):3036–3042, may 2017.

[4] P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, and I. D. Kuntz. DOCK 6: Combining techniques to model RNA-small molecule complexes. RNA, 15(6):1219–1230, apr 2009.

[5] Pedro J. Ballester and John B. O. Mitchell. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics, 26(9):1169–1175, mar 2010.

[6] Janaina Cruz Pereira, Ernesto Raúl Caffarena, and Cicero Nogueira dos Santos. Boosting docking-based virtual screening with deep learning. Journal of Chemical Information and Modeling, 56(12):2495–2506, nov 2016.

[7] Paul D Lyne. Structure-based virtual screening: an overview. Drug Discovery Today, 7(20):1047–1055, oct 2002.

[8] Adel Hamza, Ning-Ning Wei, and Chang-Guo Zhan. Ligand-based virtual screening approach using a new scoring function. Journal of Chemical Information and Modeling, 52(4):963–974, apr 2012.

[9] Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov. druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. Molecular Pharmaceutics, 14(9):3098–3104, aug 2017.

[10] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4(2):268–276, jan 2018.

[11] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. Journal of Cheminformatics, 9(1), September 2017.

[12] Daniel C. Elton, Zois Boukouvalas, Mark D. Fuge, and Peter W. Chung. Deep learning for molecular generation and optimization - a review of the state of the art. CoRR, abs/1903.04388, 2019.

[13] Tristan Aumentado-Armstrong. Latent molecular optimization for targeted therapeutic design. CoRR, abs/1809.02032, 2018.

[14] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. Journal of Chemical Information and Computer Sciences, 42(6):1273–1280, November 2002.

[15] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5):742–754, may 2010.

[16] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints, 2015.

[17] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. Journal of Computer-Aided Molecular Design, 30(8):595–608, aug 2016.

[18] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A free tool to discover chemistry for biology. Journal of Chemical Information and Modeling, 52(7):1757–1768, June 2012.

[19] Robin Winter, Floriane Montanari, Frank Noe, and Djork-Arne Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations, jul 2018.

[20] Jose Batista, Paul CD Hawkins, Robert Tolbert, and Matthew T Geballe. SiteHopper - a unique tool for binding site comparison. Journal of Cheminformatics, 6(S1), March 2014.

[21] Denis Kuzminykh, Daniil Polykovskiy, Artur Kadurin, Alexander Zhebrak, Ivan Baskov, Sergey Nikolenko, Rim Shayakhmetov, and Alex Zhavoronkov. 3d molecular representations based on the wave transform for convolutional neural networks. Molecular Pharmaceutics, 15(10):4378–4385, February 2018.

[22] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, July 2017.

[23] Nathaniel Thomas, Tess Smidt, Steven M. Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. CoRR, abs/1802.08219, 2018.

[24] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15, pages 2017–2025, Cambridge, MA, USA, 2015. MIT Press.

[25] Taco Cohen and Max Welling. Group equivariant convolutional networks. volume 48 of Proceedings of Machine Learning Research, New York, New York, USA, 20–22 Jun 2016. PMLR.

[26] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet – a deep learning architecture for molecules and materials. The Journal of Chemical Physics, 148(24):241722, June 2018.

[27] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada., pages 10402–10413, 2018.

[28] H. M. Berman. The protein data bank. Nucleic Acids Research, 28(1):235–242, jan 2000.

[29] Michael M. Mysinger, Michael Carchia, John. J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (DUD-e): Better ligands and decoys for better benchmarking. Journal of Medicinal Chemistry, 55(14):6582–6594, July 2012.

[30] Adrian M Schreyer and Tom Blundell. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. Journal of Cheminformatics, 4(1), November 2012.

[31] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained bayesian optimization for automatic chemical design. arXiv preprint arXiv:1709.05501, 2017.

[32] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman. Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. ArXiv e-prints, June 2018.

[33] Katherine Deigan Warner, Christine E. Hajdin, and Kevin M. Weeks. Principles for targeting rna with drug-like small molecules. Nature Reviews Drug Discovery, pages EP –, Jul 2018. Perspective.

[34] Esther Kellenberger, Claire Schalon, and Didier Rognan. How to measure the similarity between protein ligand-binding sites? Current Computer Aided-Drug Design, 4(3):209–220, sep 2008.

[35] Stefan Schmitt, Daniel Kuhn, and Gerhard Klebe. A new method to detect related function among proteins independent of sequence and fold homology. Journal of Molecular Biology, 323(2):387–406, oct 2002.

[36] Mu Gao and Jeffrey Skolnick. A comprehensive survey of small-molecule binding pockets in proteins. PLoS Computational Biology, 9(10):e1003302, oct 2013.

[37] Mu Gao and Jeffrey Skolnick. APoc: large-scale identification of similar protein pockets. Bioinformatics, 29(5):597–604, jan 2013.

[38] Yu-Chen Chen, Robert Tolbert, Alex M. Aronov, Georgia McGaughey, W. Patrick Walters, and Lidio Meireles. Prediction of protein pairs sharing common active ligands using protein sequence, structure, and ligand similarity. Journal of Chemical Information and Modeling, 56(9):1734–1745, sep 2016.

[39] Joshua Meyers, Nathan Brown, and Julian Blagg. Mapping the 3d structures of small molecule binding sites. Journal of Cheminformatics, 8(1), dec 2016.

[40] Daniel Reker, Anna M. Perna, Tiago Rodrigues, Petra Schneider, Michael Reutlinger, Bettina Mönch, Andreas Koeberle, Christina Lamers, Matthias Gabler, Heinrich Steinmetz, Rolf Müller, Manfred Schubert-Zsilavecz, Oliver Werz, and Gisbert Schneider. Revealing the macromolecular targets of complex natural products. Nature Chemistry, 6(12):1072–1078, November 2014.

[41] Limeng Pu, Rajiv Gandhi Govindaraj, Jeffrey Mitchell Lemoine, Hsiao-Chun Wu, and Michal Brylinski. DeepDrug3d: Classification of ligand-binding pockets in proteins with a convolutional neural network. PLOS Computational Biology, 15(2):e1006718, February 2019.

[42] Howook Hwang, Fabian Dey, Donald Petrey, and Barry Honig. Structure-based prediction of ligand–protein interactions on a genome-wide scale. Proceedings of the National Academy of Sciences, 114(52):13685–13690, December 2017.