Application of deep metric learning to molecular similarity

Damien E. Coupry* and Peter Pogány

GlaxoSmithKline, Data and Computational Sciences, UK

E-mail: damien.x.coupry@gsk.com

Abstract

Graph based methods are increasingly important in chemistry and drug discovery, with applications ranging from QSAR to molecular generation. Combining graph neural networks and deep metric learning concepts, we expose a framework for quantifying molecular similarity based on learned embeddings separate from any endpoint. Using a minimal definition of similarity, and data from the ZINC database of public compounds, this work demonstrate the properties of the embedding and its suitability for a range of applications, among them a novel reconstruction loss method for training deep molecular auto-encoders. We also compare the performance of the embedding to standard practices, with a focus on known failure points and edge cases.

Introduction

Quantifying the similarity of chemical structures has been a much used tool in drug discovery for decades,[1] and has often been adopted as a design principle for lead optimization,[2,3] under the assumption that similar molecules have a higher probability of exhibiting similar properties than dissimilar ones.[4–6] Indeed, the successful use of bioisosterism in drug development makes heavy use of the concept,[7,8] to the point that similarity is sometimes defined as a consequence of the properties, rather than the cause.[9] Most of the benchmarks for chemical structure similarity rely on this definition to compare methods,[10–12] driven in part by the availability of public activity datasets.[13] Yet, pitfalls such as so-called “activity cliffs”[14–16] should moderate the confidence in the underlying principle. Furthermore, other use cases of similarity exist, and are not captured by the similar properties paradigm: patent mining and infringement prediction,[17] building block selection for synthesis, retrosynthesis and scaffold hopping,[18–20] molecular generation evaluation,[21] etc. A “good” measure of similarity should ideally show equal performance in all these applications, never relying too much on any one definition or type of benchmark.

On the practical side, similarity can be more generally understood as the combination of a molecular representation and an appropriate metric.[3] Today, the combination of two-dimensional molecular circular fingerprints[22,23] with the Tanimoto coefficient[24] is still the most widely used, and generally hard to outperform in traditional benchmarks.[25] Still, these methods suffer from a number of identified drawbacks, regularly analysed but difficult to route around in the absence of a more general representation.[26,27] Most of the recent efforts to develop original molecular encodings focus on the relational nature of molecules as seen in a 2D context. By considering structures as a graph with atoms as nodes and bonds as edges, we can draw on the considerable field of extant work on graph similarity in general: computationally expensive graph edit distance, graph isomorphism quantification or maximum common subgraph,[28–32] graph kernels for similarity,[33] and the increasingly popular deep learning algorithms.[34] The latter rely on embeddings learned from variational reconstruction tasks,[35] end-to-end property predictions,[36] or borrow architectures from facial recognition.[37]

In this work, we leverage the ability of graph neural networks from the Deep Graph Library[38,39] to learn chemical structures embeddings using the triplet loss,[40] to our knowledge the first such use of it. A training dataset is constructed automatically using a minimal definition of molecular similarity and public compounds. We show that these embeddings satisfy the conditions to be considered an improved encoding of chemical information in both traditional benchmarks and novel applications.

Experiments

Dataset generation

The ZINC database was downloaded (1.487 billion compounds)[41,42] and processed as follows. Parent structures were created, bad valencies, compounds with poorly defined bonds, isotope labelled compounds and compounds containing elements other than N, O, C, S, F, Cl, Br and I were removed. This initial filtering removed around 2 million compounds. Reduced Graphs,[43,44] Bemis-Murcko graph and detailed frames[45] were generated for each compound. In the Reduced Graph, the full molecular graph is reduced to pharmacophore feature type nodes. Whereas the Bemis-Murcko graph frames contain the anonymous frame of the molecule without the side chains, atom types and bond orders. The Bemis-Murcko detailed frame contains the frame of the molecule (side chains removed) with atom types and bonds marked. Comparison of these molecular representations is given on Figure 1.

REOS[46] and PAINS A[47] filters were applied on the remaining compounds and molecular weight (MW) was calculated to remove everything with MW>650 daltons, thus keeping 1.199 B compounds. Compounds were clustered in three ways:

1. Having the same Reduced Graph and Graph Frame (GFRG)
2. Having the same Reduced Graph and Detailed Frame (DFRG)
3. Having the same Reduced Graph (RG)

Most of the processing after this was done using BIOVIA Pipeline Pilot.[48] All compounds belonging to a GFRG cluster with less than 4 members were removed. In the case of compounds belonging to GFRG clusters with more than 10k members, DFRG clusters were used

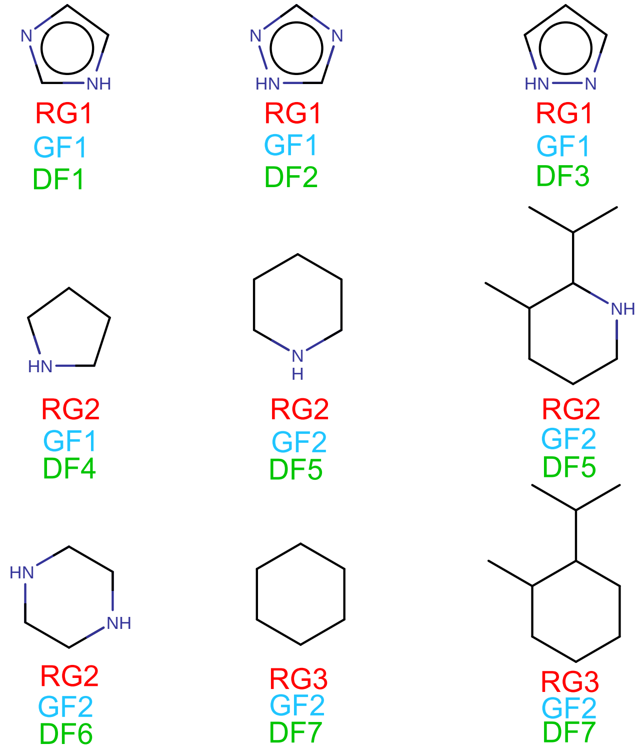

Figure 1: A comparison of the Reduced Graph (RG), Bemis-Murcko graph (GF) and detailed frames (DF) clusters. The numbers after the character show the cluster. RG1 is a cluster of aromatic ring containing compound which contain hydrogen bond donor and acceptor. RG2 are aliphatic rings with hydrogen bond donors, RG3 are aliphatic rings without feature. There are only two graph frame clusters: 5-membered rings (GF1) and 6-membered rings (GF2). Detailed frames are only identical, if the compounds differ in ring substituents connected to rings with single bonds (DF5 and DF7).

in place of GFRG. For DFRG clusters, a maximum size of 20k members was established, with random subsampling performed on clusters above this limit. 1.13 billion compounds remained and cluster centers were assigned to them. Cluster Molecules component of BIOVIA Pipeline Pilot[48] was used to determine the cluster centroids for each cluster defined above (ECFP4 and heavy atom count was used for getting the centroids). For every cluster the number of identities was calculated. If the number of identities was larger than 0.4, all the cluster elements were discarded. 1.113 billion compounds remaind in 16.71 million clusters. The number of clusters for each Reduced Graph was calculated and only Reduced Graphs which have at least 2 clusters were kept (1.059 billion compounds).

The triplet loss trains networks by contrasting a reference structure with two additional compounds, called positive and negative controls. The positive control should be qualitatively similar to the reference. For this purpose, the two were selected randomly from within the same cluster (GFRG cluster for the initial smaller clusters, for the larger clusters, where GFRF cluster size $\geq$ 10,000, DFRG clusters are used). The negative control should conversely be less similar to the reference than the positive. Selecting a very different compound is not optimal, since the chemical space size increase towards larger dissimilarities. Thus, while it would be correct to choose a negative control from a different cluster, choosing a compound that has *some* similar features to the reference is more valuable to the training process. Therefore we have randomly selected the negative control from a different cluster than the cluster of the reference, but their Reduced Graph should be the same. This way 12’361’633 triplets were created. A detailed schema of the data preparation can be seen on Figure 2.

Model training

For all training and benchmarking purposes, the random seed is fixed at 42 for repeatability, and the hyperparameters have been kept unoptimized and to the default values to prevent bias. We used the DGL-Lifesci open source framework for computations on graphs, and its

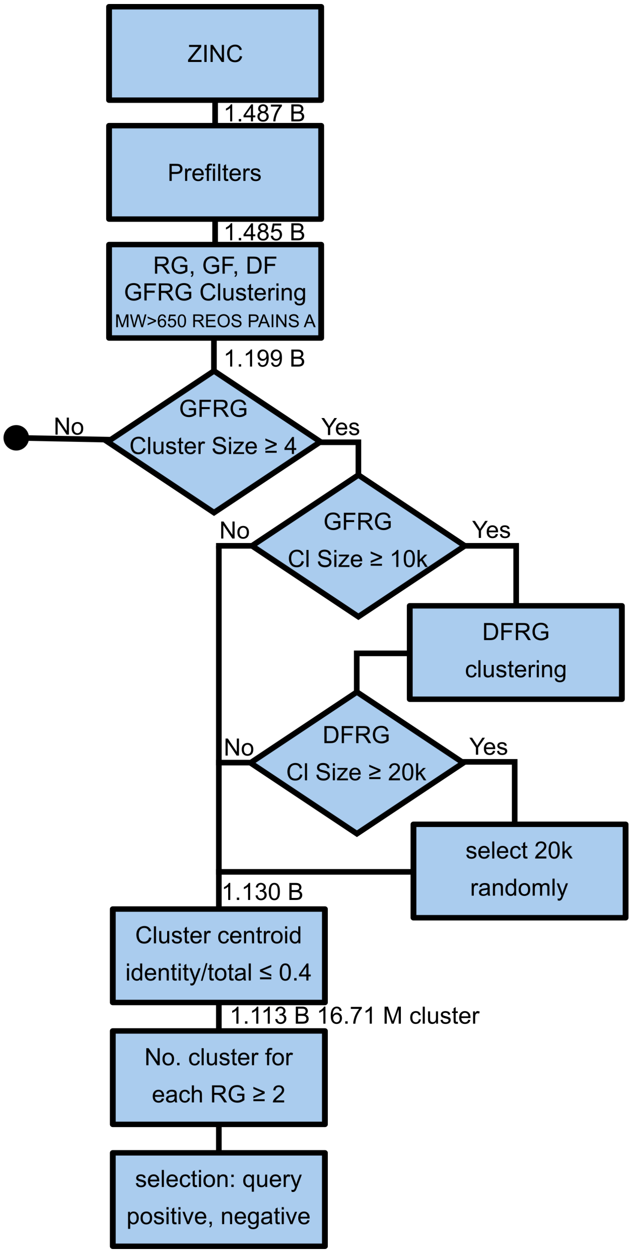

Figure 2: The process diagram of data preparation.

message passing neural network implementation (MPNNPredictor)[49] as model architecture. This type of model repeatedly accumulates bond information as well as node information based on connectivity, and has been used with great effect in state of the art QSAR applications.[50] We chose to use the default parameters and an output size equal to 16 as an embedding dimension (*n_tasks*). The input for such a model are molecular graphs, which are obtained using the CanonicalAtomFeaturizer and CanonicalBondFeaturizer from DGL. The details of what is included in the graphs features can be found in the DGL-lifesci documentation.[51,52] These representations are regularized with a node ablation probability of 1% and edge ablation probability of 5%. At each step of the training, an instance of the MPNN is used to embed each of the three graphs of the input (anchor, positive and negative); the triplet margin loss from pytorch[53] then updates the weights of the network to maximize the distance between the anchor and negative, while minimizing the distance between the anchor and the positive, as seen in Figure 3.

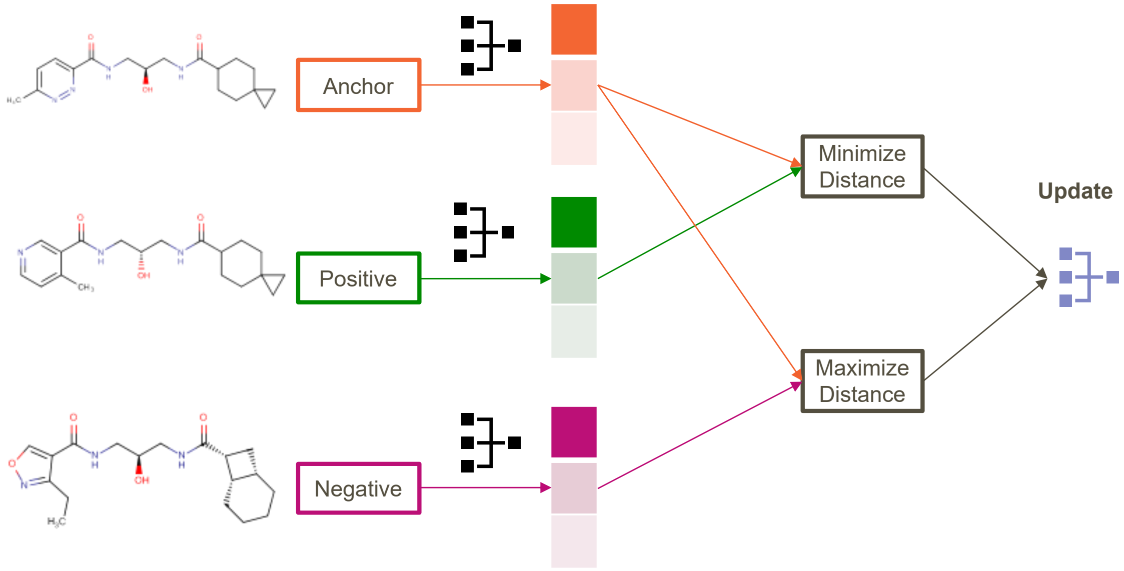

Figure 3: The architecture of the triplet loss embedding during training.

The training used the pytorch-lightning framework[54] with a 25 epochs early stopping criterion, the Adam optimizer with the default learning rate of 10.0^{-3}, and took two days on an Nvidia GEFORCE1080 GPU with a batch size of 128. For more details, hyperparameters, and training curves, please refer to the project's github page.

Benchmarks choice

The benchmarks for the present use case should optimally measure a number of things:

- The performance on popular applications; here the activity classification tasks such as the ones described in Riniker *et al.*[12]

- The performance on edge cases, such as the ones described in Flower *et al*,[26] particularly when the failure of traditional fingerprint based similarity measure is due to the basic technique of fragmentation.

- The condition of graph isomorphism: the ordering of the molecule atoms and bonds should have no influence on the embedding.

Additionally, *desired* properties of an encoding come from the coupling with a metric. In particular, using a euclidean distance metric on a well defined euclidean vector space gives rise to a number of interesting properties:

- very fast querying and operations

- Similarity can be defined with respect to geometric elements: around a barycentre, along a path between molecules, within a cone, etc.

- the space and metric together are unbound in value for dissimilarity: there are many more ways of being dissimilar than similar, and the distances distribution could reflect that.

Results

Activity prediction tasks benchmarking

While an imperfect measure of fitness for any new chemical embedding, the dominance of benchmarking platforms making use of a variety of activity prediction datasets makes it an obligatory step in evaluating any new contribution. In particular, it enables two separate conclusions to be reached:

1. Whether the information contained in the embedding is sufficient to fit models successfully, regardless of compared performance

2. Whether these models are statistically different from references to demonstrate the originality of the embedding

To answer the second query, it is necessary to benchmark models on a suitably high number of instances for each class. For this purpose, a dataset of IC50 activities was extracted from the ChEMBL28 database. All targets with a unique structure count between 5k and 20k were kept, with activity threshold automatically set at the 75th percentile of the PIC50 values if and only if this is superior by at least one standard deviation from the minimum value and maximum value. This classification task was modelled by a k-nearest neighbours classifier from the scikit-learn python package,[55] trained on ECFP0 and ECFP4 fingerprints from the rdkit package,[56] as well as on learned embeddings . Only targets with an ECFP0 5-fold stratified cross validation Cohen's Kappa score above 0.25 were kept, to constrain the benchmark tasks to be relatively hard but tractable, resulting in a set of 55 targets. For each triplet of models, the Cochran's Q test was applied to verify statistical difference. The p-values of 30 tested targets were <0.05 and sufficient to reject the null hypothesis that all the models were equivalent. Subsequent confirmation with pairwise McNemar tests with Bonferroni correction show the embedding models to be the source of the statistical difference, thus answering our second point. The performances on this final set of 30 targets are shown in Figure 4, and answers our first point to our satisfaction.

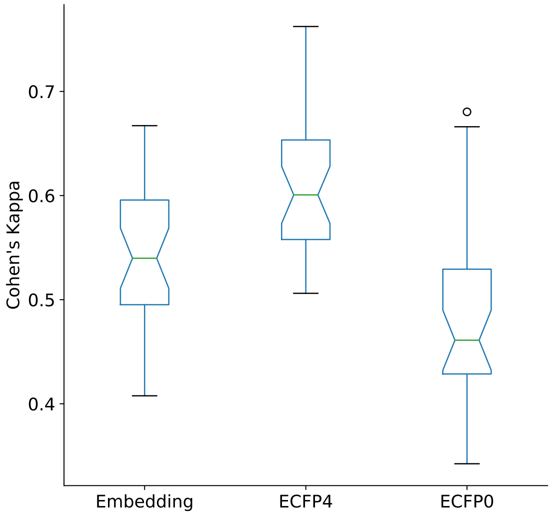

Figure 4: Performance in activity classification tasks from ChEMBL28.

Failure points of circular fingerprints

One noted effect of the bit-string fingerprints is the skewing effect of size on the distribution of similarities as illustrated in Figure 6 of Flowers *et al.*[26] Applying the same reference set of compounds for comparison on a diverse set of molecules using the MPNN learned embedding leads to a much better shape of the distributions. While the larger molecule has a more chaotic profile of similarity (probably due to the fact that the larger a structure, the more ways for something to be similar to it), it otherwise seems independent from the size of the molecules. This is shown in Figure 5.

Another point where fingerprints fail to accurately describe molecular similarity is the case of molecules with repeated motifs. When using Tanimoto similarity of circular fingerprints in bit string form, the similarity tapers off quickly to a fixed non-zero value. The learned embedding is immune to this effect. Likewise, the insertion of moieties within a scaffold has an unduly small effect when it does not perturb the fragmentation of the structure

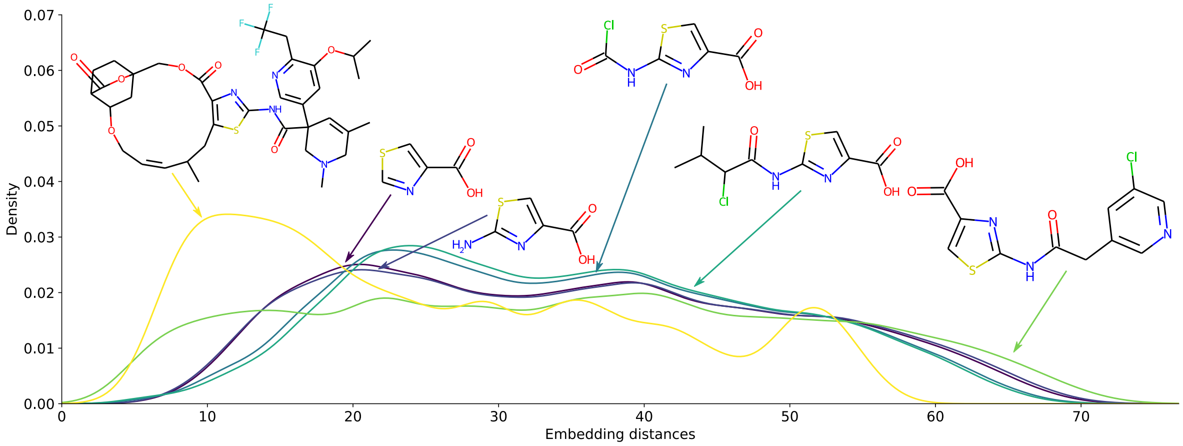

Figure 5: Distribution of embedding distances of 5 references compounds to a diverse set of 120k compounds from the Zinc database.

by fingerprints, but is correctly shown to matter a lot by the embedding. In addition, it also retains the concepts of fragments, aromaticity, and some level of isosterism. Some examples illustrating these points are shown in Figure 6.

Additional properties

As stipulated earlier, the distribution of similarities should be notably different between positive examples and negative examples: the first distribution should show a sharp peak around optimal similarity, and the second should display a long tail representing the many different sources of dissimilarity. After applying both the ECFP4 Tanimoto coefficient comparison and the learned MPNN embedding to unseen triplets of our generated dataset, we indeed see such a behaviour illustrated in 7.

Another critical desired property for a novel molecular distance measure is the ability to correctly compare partial and *chemically invalid* molecular graphs and provide gradient information. This leads to the important fact that trained embeddings are essentially derivable reconstruction loss with a quadratic energy surface, with widespread potential applications. For example:

- Accelerated training of reconstruction based molecular generators such as variational

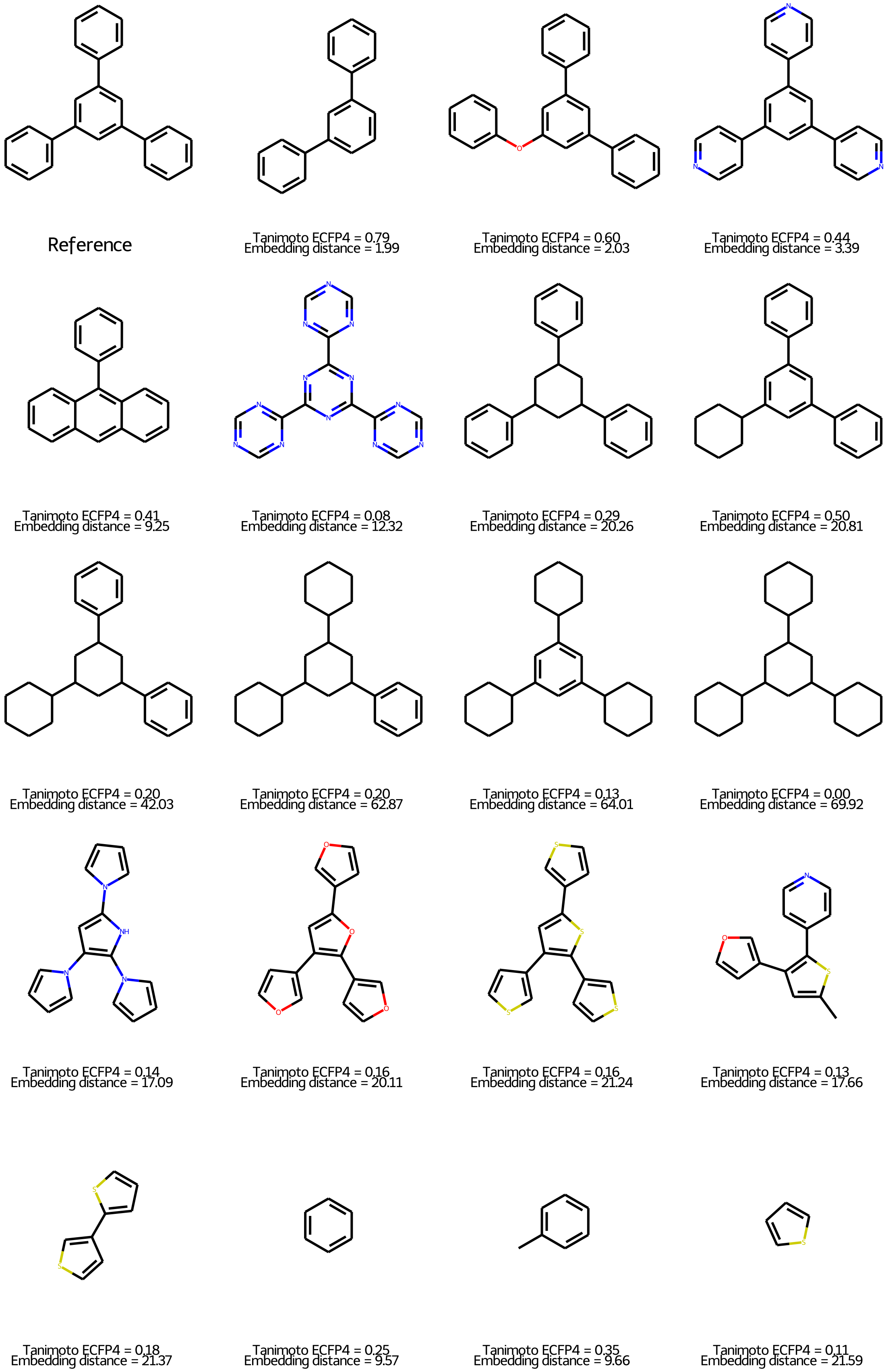
Reference
Tanimoto ECFP4 = 0.79
Embedding distance = 1.99
Tanimoto ECFP4 = 0.60
Embedding distance = 2.03
Tanimoto ECFP4 = 0.44
Embedding distance = 3.39
Tanimoto ECFP4 = 0.41
Embedding distance = 9.25
Tanimoto ECFP4 = 0.08
Embedding distance = 12.32
Tanimoto ECFP4 = 0.29
Embedding distance = 20.26
Tanimoto ECFP4 = 0.50
Embedding distance = 20.81
Tanimoto ECFP4 = 0.20
Embedding distance = 42.03
Tanimoto ECFP4 = 0.20
Embedding distance = 62.87
Tanimoto ECFP4 = 0.13
Embedding distance = 64.01
Tanimoto ECFP4 = 0.00
Embedding distance = 69.92
Tanimoto ECFP4 = 0.14
Embedding distance = 17.09
Tanimoto ECFP4 = 0.16
Embedding distance = 20.11
Tanimoto ECFP4 = 0.16
Embedding distance = 21.24
Tanimoto ECFP4 = 0.13
Embedding distance = 17.66
Tanimoto ECFP4 = 0.18
Embedding distance = 21.37
Tanimoto ECFP4 = 0.25
Embedding distance = 9.57
Tanimoto ECFP4 = 0.35
Embedding distance = 9.66
Tanimoto ECFP4 = 0.11
Embedding distance = 21.59

Figure 6: Selection of pairwise comparisons illustrating a diverse set of molecular similarities.

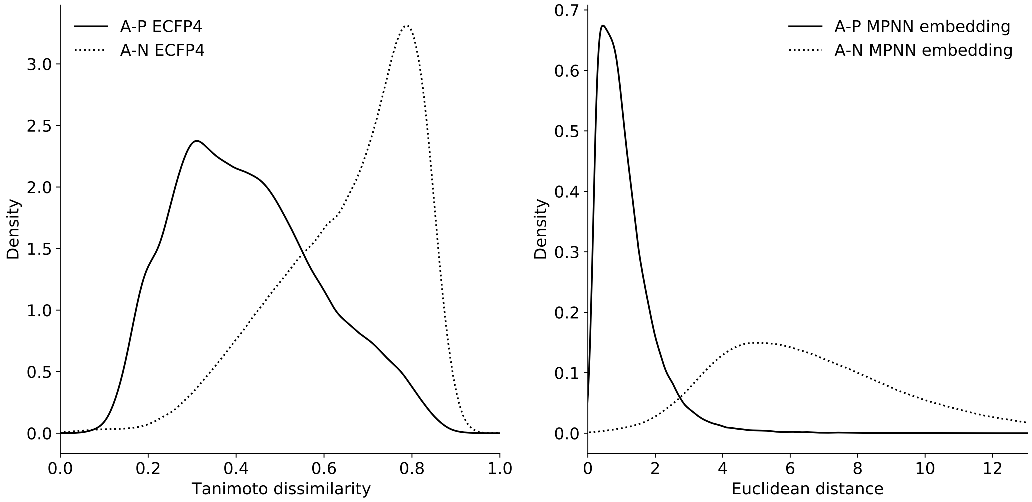

Figure 7: Comparison of the similarity distributions on unseen triplets

auto-encoders.

- Additional information in tasks such as missing edge and node prediction.

- Chemical subspace constraints for conditional molecular generators

These tasks are deeply unsuitable to traditional fingerprints or property based similarity : for most of the training process, the molecular graphs on which computation happens are completely invalid, the chemical information on what is a molecule still being accrued. Yet a learned embedding, as is shown in Figure 8, is very robust to node and edge deletion, demonstrating a quasi linear distance relationship with the number of deleted elements. This is an exciting property, and we look forward to seeing it explored further.

Finally, a critical property of the embedding is its ability to be used in conjunction with transfer learning,[57,58] and be retrained on particular subsets of the chemical space according to tailored similarities obtained from SAR, Molecular Matched Pairs,[59] or a more complex multiple-parameters function. Such a retrained model would retain the general concepts of molecular graph similarity while quickly converging to a more appropriate representation of the problem at hand, thus sparing resources in training and data gathering.

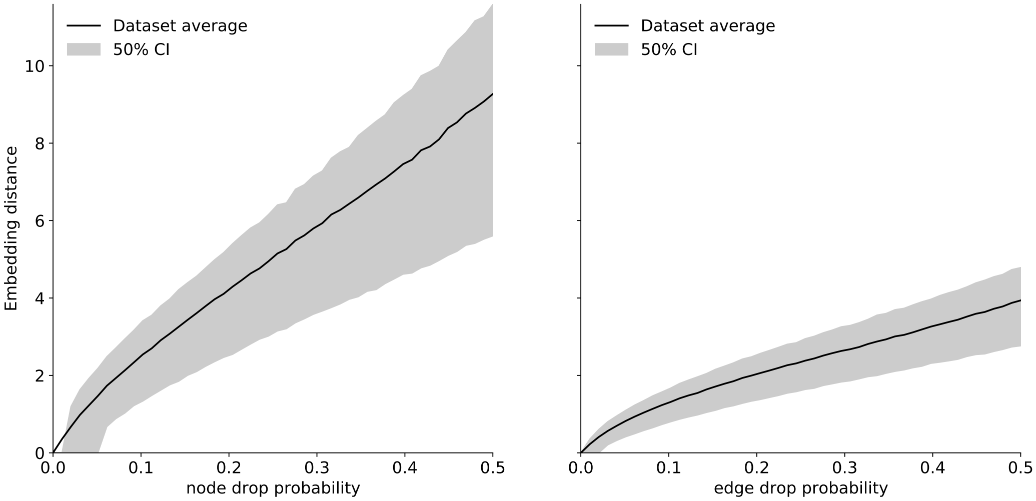

Figure 8: Effect of random element deletion on embedding distance. No comparison with ECFP4 could be obtained due to the overwhelming rate of invalidity of the resulting structures.

Conclusions

We have shown that using the triplet margin loss jointly with molecular graph based deep neural networks trains latent representations that satisfy the many definitions of chemical similarity. A naive example of such an embedding was trained with no hyperparameters optimization on a dataset constructed from public molecules and some basic concepts of graph similarity. This naive example compares acceptably out of the box with the accepted standard of circular fingerprints Tanimoto scores, while possessing many additional properties such as being derivable or retrainable. We believe such properties may be of great use to train reconstruction based molecular generators.

All code and data is available on github.com and is sufficient to reproduce our conclusions and graphs.

The authors were both employed by GlaxoSmithKline UK at the time of the present work.

Acknowledgement

The authors thanks Darren Green and Kim Branson for their preliminary review; as well as the entire Molecular Design team for their constructive feedback.

Supporting Information Available

The following files are available free of charge.

- lightning_logs.zip: a compressed file containing the detailed training logs of the example model presented in this work. Visualize using the tensorboard utility, as described here: pytorch-lightning.readthedocs.io
- model.pt: the weights and hyperparameters checkpoint of the example model presented in this work. Should be used in conjunction with the code released on github.

References

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *Journal of chemical information and computer sciences* **1998**, *38*, 983–996.

(2) Kubinyi, H. Similarity and dissimilarity: a medicinal chemist's view. *Perspectives in Drug Discovery and Design* **1998**, *9*, 225–252.

(3) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2014**, *57*, 3186–3204.

(4) Johnson, M. A.; Maggiora, G. M. *Concepts and applications of molecular similarity*; Wiley, 1990.

(5) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E.

Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *Journal of medicinal chemistry* **1996**, *39*, 3049–3059.

(6) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry* **2002**, *45*, 4350–4358.

(7) Patani, G. A.; LaVoie, E. J. Bioisosterism: a rational approach in drug design. *Chemical reviews* **1996**, *96*, 3147–3176.

(8) Lima, L. M.; Barreiro, E. J. Bioisosterism: a useful strategy for molecular modification and drug design. *Current medicinal chemistry* **2005**, *12*, 23–49.

(9) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry* **2004**, *2*, 3204–3218.

(10) Irwin, J. J. Community benchmarks for virtual screening. *Journal of computer-aided molecular design* **2008**, *22*, 193–199.

(11) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *Journal of chemical information and modeling* **2009**, *49*, 169–184.

(12) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics* **2013**, *5*, 1–17.

(13) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B., et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012**, *40*, D1100–D1107.

(14) Maggiora, G. M. On outliers and activity cliffs why QSAR often disappoints. 2006.

(15) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2012**, *55*, 2932–2942.

(16) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent progress in understanding activity cliffs and their utility in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2014**, *57*, 18–28.

(17) Rhodes, J.; Boyer, S.; Kreulen, J.; Chen, Y.; Ordonez, P. *Biocomputing 2007*; World Scientific, 2007; pp 304–315.

(18) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold hopping. *Drug discovery today: Technologies* **2004**, *1*, 217–224.

(19) Boehm, M.; Wu, T.-Y.; Claussen, H.; Lemmen, C. Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *Journal of medicinal chemistry* **2008**, *51*, 2468–2480.

(20) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **2017**, *3*, 1237–1245.

(21) Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications* **2020**, *11*, 1–10.

(22) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.

(23) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

(24) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **2015**, *7*, 1–13.

(25) Raymond, J. W.; Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of computer-aided molecular design* **2002**, *16*, 59–71.

(26) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *Journal of chemical information and computer sciences* **1998**, *38*, 379–386.

(27) Dixon, S. L.; Koehler, R. T. The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *Journal of medicinal chemistry* **1999**, *42*, 2887–2900.

(28) Garcia-Hernandez, C.; Fernández, A.; Serratosa, F. Ligand-based virtual screening using graph edit distance as molecular similarity measure. *Journal of chemical information and modeling* **2019**, *59*, 1410–1421.

(29) Bunke, H.; Shearer, K. A graph distance metric based on the maximal common subgraph. *Pattern recognition letters* **1998**, *19*, 255–259.

(30) Bunke, H.; Allermann, G. Inexact graph matching for structural pattern recognition. *Pattern Recognition Letters* **1983**, *1*, 245–253.

(31) Dijkman, R.; Dumas, M.; García-Bañuelos, L. Graph matching algorithms for business process model similarity search. International conference on business process management. 2009; pp 48–63.

(32) Berretti, S.; Del Bimbo, A.; Vicario, E. Efficient matching and indexing of graph models in content-based retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2001**, *23*, 1089–1105.

(33) Kriege, N. M.; Johansson, F. D.; Morris, C. A survey on graph kernels. *Applied Network Science* **2020**, *5*, 1–42.

(34) Ma, G.; Ahmed, N. K.; Willke, T. L.; Philip, S. Y. Deep graph similarity learning: A survey. *Data Mining and Knowledge Discovery* **2021**, 1–38.

(35) Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. International conference on machine learning. 2018; pp 2323–2332.

(36) Brown, N. Chemoinformaticsan introduction for computer scientists. *ACM Computing Surveys (CSUR)* **2009**, *41*, 1–38.

(37) Bai, Y.; Ding, H.; Bian, S.; Chen, T.; Sun, Y.; Wang, W. Simgnn: A neural network approach to fast graph similarity computation. Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. 2019; pp 384–392.

(38) Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y., et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315* **2019**,

(39) Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; Karypis, G. DGL-LifeSci: An Open-Source Toolkit for Deep Learning on Graphs in Life Science. *arXiv preprint arXiv:2106.14232* **2021**,

(40) Schultz, M.; Joachims, T. Learning a distance metric from relative comparisons. *Advances in neural information processing systems* **2004**, *16*, 41–48.

(41) Sterling, T.; Irwin, J. J. ZINC 15 Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337, PMID: 26479676.

(42) ZINC database. 2021; `http://files.docking.org/2D/`, Last accessed 27/01/2021.

(43) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 338–345, PMID: 12653495.

(44) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 2145–2156, PMID: 15554685.

(45) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893, PMID: 8709122.

(46) Walters, W.; Stahl, M. T.; Murcko, M. A. Virtual screeningan overview. *Drug Discovery Today* **1998**, *3*, 160–178.

(47) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry* **2010**, *53*, 2719–2740, PMID: 20131845.

(48) BIOVIA, D. S. Discovery studio visualizer, Release 2020, San Diego: Dassault Systèmes, 2019. 2020; `http://accelrys.com/products/collaborative-science/biovia-discovery-studio/visualization-download.php`, Last accessed 07/07/2021.

(49) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. International conference on machine learning. 2017; pp 1263–1272.

(50) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **2019**, *59*, 3370–3388.

(51) dgllife.utils.CanonicalAtomFeaturizer. `https://lifesci.dgl.ai/generated/dgllife.utils.CanonicalAtomFeaturizer.html#dgllife.utils.CanonicalAtomFeaturizer`, Accessed: 2021-07-01.

(52) dgllife.utils.CanonicalBondFeaturizer. `https://lifesci.dgl.ai/generated/dgllife.utils.CanonicalBondFeaturizer.html`, Accessed: 2021-07-01.

(53) Paszke, A. et al. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 8024–8035.

(54) Falcon, e. a., WA PyTorch Lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning* **2019**, *3*.

(55) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(56) Landrum, G. RDKit: Open-Source Cheminformatics Software. **2021**,

(57) Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. International conference on artificial neural networks. 2018; pp 270–279.

(58) Weiss, K.; Khoshgoftaar, T. M.; Wang, D. A survey of transfer learning. *Journal of Big data* **2016**, *3*, 1–40.

(59) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool: miniperspective. *Journal of medicinal chemistry* **2011**, *54*, 7739–7750.