Supporting Information: Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds

Qiyuan Zhao, Nicolae C. Iovanac, and Brett M. Savoie*

Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, 47906

E-mail: bsavoie@purdue.edu

Implementation details for neural fingerprint model

We examined a number of different molecular representations (including one-hot encodings, Morgan fingerprints, and latent vectors produced by autoencoder models[1]) and model types (including feed forward, convolutional, and reccurrent neural networks) in attempting to predict the ring correction. The best performance was achieved through the use of learned neural molecular fingerprints[2] fed to a multi-layer perceptron. Rather than relying on a predefined function to create the molecular fingerprint, it is instead learned in tandem with the training of the property prediction model. This ensures that the features represented in the fingerprint vector are those best suited for the task of predicting the ring effect on the enthalpy of formation. The neural fingerprint network itself consists of a series of graph convolutions that operate over atoms and their neighboring features to develop a real-valued vector of learned molecular features.

Taking inspiration from the physical nature of our problem, that is, predicting the difference in enthalpy between a cyclic compound when considered up to a depth of 2 bonds from

the central ring and its unsubstituted (or depth of 0) form, we tested and confirmed that the best results were achieved by providing *both* the SMILES string of the given molecule as well as the SMILES string of its base ring (*e.g.* for ethylbenzene the SMILES string corresponding to benzene is also provided as input). The two SMILES strings are fed as inputs to two separate neural fingerprint stacks, (note the SMILES strings define the molecular graphs) the outputs of which are then combined to form a "composite" neural molecular fingerprint, where the features of the ring are highlighted against the features of the given molecule as a whole. This composite fingerprint is then fed as input to a multilayer perceptron terminating in a single linear unit which produces the ring correction (i.e., difference between the linear TCIT prediction and the labeled value) to the enthalpy of formation as output. A schematic of the model is provided in Figure S1. Hyperparameters, including the learning

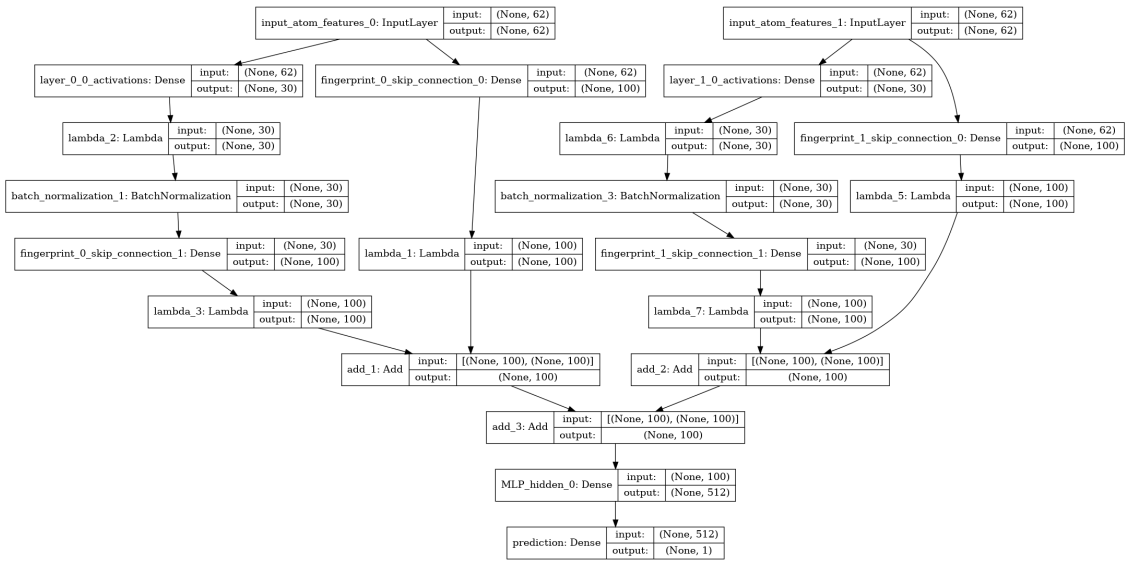

Figure S1: Architecture for neural fingerprint-based ring correction model. Note that two identical fingerprint models are trained in tandem to learn features of the input molecule and its base cyclic structure. As the size of the neural fingerprint and the depth of the MLP was adjusted during hyperparameter optimzation, the architecture of the model used to directly predict enthalpy of formation is slightly different.

rate, size of the neural fingerprint, molecular convolutional width, and the size and depth of the subsequent multi-layer perceptron, were evaluated and adjusted according to a grid based search. Models were constructed using the Keras API[3] with Tensorflow backend.[4] We

based our approach on the Keras-implementation of the neural fingerprint model described by Duvenaud *et al.*: https://github.com/GUR9000/KerasNeuralFingerprint.[2]

Performance of purely machine learning approaches

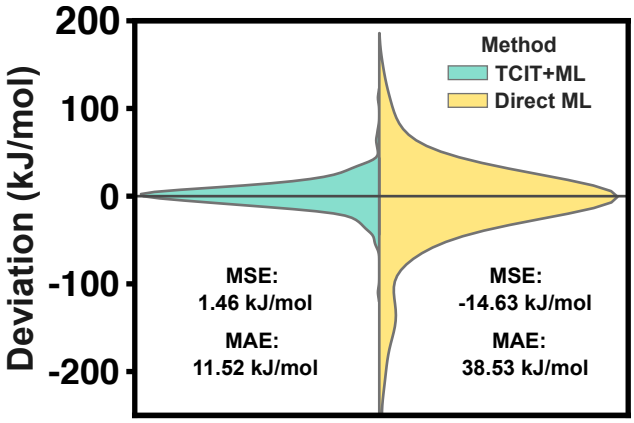

Figure S2: Comparison of TCIT methodology with ring corrections evaluated by transfer learning model versus direct enthalpy of formation prediction with neural fingerprint model.

Due to the success of the transfer learning model in evaluating the ring correction to the enthalpy of formation, we also tested neural fingerprint models to directly predict the enthalpy of formation (ΔH_{f}). Two models were trained on the same training set prepared for the transfer learning model and were tested on the ring model compounds obtained from the PNK database. One model is designed to directly predict the ring correction and the other is trained to predict ΔH_{f}. The mean absolute error (MAE) for these two models are 7 kcal/mol and 7.7 kcal/mol, respectively, which are much larger than the MAE of the TCIT+ML scheme. The direct ΔH_{f} prediction model was also applied to the original PNK dataset, which contains some molecules larger than depth=2 rings, and the resulting comparison is shown in Figure. S2. The MSE and MAE of the direct ML model are both much larger than TCIT+ML model which suggests that the hybridization of ML and physics-based models provides more accurate and reliable predictions, especially when the training data is not sufficiently large.

Lists of excluded compounds, experimental values and G4 calculation results of model compounds

Table S1: List of 10 excluded model compounds and their ΔH_f (kJ/mol).

Name	Exp	G4
bromopentafluorobenzene	-712.00	-761.02
cis-4-methylcyclohexanol	-347.50	-321.26
trans-2-methylcyclohexanol	-352.50	-320.30
cis-3-methylcyclohexanol	-350.90	-321.30
trans-4-methylcyclohexanol	-367.20	-321.26
1,3-dioxocane	-336.80	-373.33
3-chlorophenol	-153.30	-124.52
4-chlorophenol	-145.80	-122.14
3-methyl-2,5-furandione	-504.50	-452.20
1,7-bis(methylamino)-1,3,5-cycloheptatriene	211.10	194.14

Table S2: List of 22 excluded intermediate sized compounds and their ΔH_f (kJ/mol).

Name	Exp	TCIT-RC1
1-methyl-4-(1-butenyl-sulphonyl)-benzene	-229.80	-197.653987
diphenylethanedione	-55.50	-21.267968
dibenzoyl-peroxide	-271.70	-224.437079
1,2-diphenylethanone	22.30	49.438074
1,4-benzenedicarboxamide	-292.00	-262.037331
benzoic-anhydride	-319.00	-281.606077
1-methyl-4-(1,2-propadienylsulphonyl)-benzene	-32.60	2.086912
1,4-benzenedicarboxylic-acid	-717.90	-648.462880

(ethenylsulphonyl)-benzene	-129.00	-106.468591
4-methoxybenzaldehyde	-202.70	-160.002927
phenyl-2-hydroxybenzoate	-344.50	-279.912008
1,3-benzenedicarboxamide	-294.00	-262.667450
carbonic-acid-diphenyl-ester	-311.20	-256.094604
cis-azobenzene	458.10	371.029437
n-phenylacetamide	-128.90	-81.364235
1,3-benzenedicarboxylic-acid	-696.30	-649.092999
tetraphenylmethane	397.80	481.550414
1,3,5-triphenylbenzene	367.50	440.349331
2,4,6-triisopropyl-benzophenone	-188.50	-146.794671
diphenylether	52.00	96.202721
1,1-diphenylethylene	245.60	273.583227
acetic-acid-3-methylphenyl-ester	-313.40	-287.451727

Table S3: Model compounds with Experimental and G4 ΔH_f (kJ/mol).

Name	Exp	G4
3,5-dimethylphenol	-161.50	-159.39
3,5-dimethylpyridine	72.80	70.28
pentachlorophenol	-225.10	-208.46
cis-2-methylcyclohexanol	-327.00	-320.30
methylenecyclopentane	12.00	12.24
1-amino-2,4,6-cyclohepta-trien-1-one	39.50	31.69
pyrimidine	195.90	185.41
1,3,5,7-cyclooctatetraene	295.90	298.69
2,3-dihydrothiophene	90.70	80.12

3-pyridinol	-43.70	-31.66
1-ethyl-2-nitrobenzene	11.20	9.98
2,3,5-trichloro-2,5-cyclohexadiene-1,4-dione	-180.70	-185.51
bromobenzene	105.40	100.59
1,3,5,7-tetroxane	-620.20	-628.97
1,2-dimethylbenzene	19.10	16.14
1,3-dinitrobenzene	53.80	42.63
trans-3-methylcyclohexanol	-329.10	-321.30
2,3-dimethylpyridine	68.30	63.25
2-methyl-6-hydroxypyridine	-120.20	-116.27
4-methylpyridine	104.10	102.30
3-methylcyclopentene	7.40	7.58
trans-1,2-dimethylcyclohexane	-179.90	-174.52
isoxazole	78.60	79.87
thiepane	-65.80	-65.45
nitrobenzene	67.50	56.46
oxazole	-15.50	-16.59
cyclopentanone	-192.10	-192.85
chlorobenzene	52.00	50.30
2-pyridinol	-79.70	-73.96
1,3-dichlorobenzene	25.70	33.58
cyclodecane	-154.30	-151.09
1,3,5-triazine-2,4,6-triamine	51.80	62.00
1,3-dioxolane	-298.00	-299.86
4-methylene-2-oxetanone	-190.20	-192.87
thietane	60.60	61.24

cyclohexylamine	-104.90	-101.99
oxetane	-80.50	-80.27
3-methylisoxazole	38.60	37.03
methyloxirane	-94.70	-95.44
cyclopentene	33.90	37.41
1-hydroxy-2,2,6,6-tetra-methyl-4-piperidinone	-298.00	-325.02
pyridazine	278.30	278.42
cyclopentane	-76.40	-72.73
3,3-dimethyloxetane	-148.20	-148.72
3,4-dimethylpyridine	70.10	66.72
1,4-benzenedicarboxamide	-292.00	-279.23
1,3,5-cycloheptatriene	180.90	184.95
1,3-dimethylbenzene	17.30	16.93
2,5-dimethyl-1h-pyrrole	39.80	37.50
2,4-dimethylpyridine	63.90	60.63
tetrahydro-2h-thiopyran	-63.50	-64.59
2,4-dimethylphenol	-162.90	-158.03
fluorobenzene	-116.00	-112.59
dihydro-3-(2h)-thiophenone	-135.30	-136.85
1,3,5-trioxane	-465.90	-470.88
2h-thiete-1,1-dioxide	-124.60	-124.19
2,6-dimethylphenol	-161.80	-162.85
cyclobutane-1,3-dione	-186.30	-182.59
2,5-dihydro-3-methyl-thiophene-1,1-dioxide	-291.90	-273.42
1,2-difluorobenzene	-293.80	-289.79
cyclopropane	53.30	54.17

2-methyl-5-hydroxypyridine	-69.80	-71.24
1,3-cyclopentadiene	134.30	135.33
trans-1,2-dimethyl-cyclopentane	-136.60	-136.30
1,3-benzenediol	-274.70	-268.36
1,4-dimethylbenzene	18.00	17.56
2,3-dihydrothiophene-1,1-dioxide	-262.00	-225.25
3-methylphenol	-132.30	-125.82
pentafluorophenol	-956.80	-954.60
methylenecyclohexane	-25.20	-33.22
cyclobutane	28.40	28.39
tetrahydrothiophene	-34.10	-33.76
(z,z)-1,5-cyclooctadiene	101.10	101.87
tetrahydro-2-methyl-thiophene	-64.20	-65.84
pyrazine	196.10	203.89
dihydro-2-(3h)-thiophenone	-196.20	-173.24
methylenecyclopropane	200.50	192.46
2,5-dihydrothiophene	86.90	82.49
2-methylpyridine	99.20	97.41
3,4-dimethylphenol	-156.60	-157.61
2-methylphenol	-128.60	-127.93
cyclohexanone	-226.10	-228.39
3-methylthiophene	82.60	77.21
2,5-cyclohexadiene-1,4-dione	-122.90	-117.20
pentafluorobenzene	-806.50	-802.85
trimethylthiirane	-21.50	-30.95
3-amino-5-methylisoxazole	19.70	31.10

2,2-dimethylthiirane	11.30	0.50
cyclopropylamine	77.00	76.78
cis-1,2-dimethyl-cyclopentane	-129.50	-136.30
2,4,6-cycloheptatrien-1-one	43.90	63.52
1,3-dithiole-2-thione	253.00	195.84
2-methyl-2-oxazoline	-130.80	-126.90
hexamethylbenzene	-86.80	-79.35
2,5-dimethylpyridine	66.50	63.61
tetrahydro-2h-pyran	-223.40	-221.99
2,4,6-trimethyl-1,3,5-trioxane	-631.80	-625.00
4-methyl-1,3-dioxolan-2-one	-563.70	-555.98
benzene	82.60	83.53
hexafluorobenzene	-955.40	-953.08
pyridine	140.40	139.08
1-methyl-4-nitrobenzene	31.00	20.58
furan	-34.90	-34.47
pyrrolidine	-3.40	-1.77
2,6-dichloro-2,5-cyclohexa-diene-1,4-dione	-174.50	-165.08
2,5-dihydro-2-methyl-thiophene-1,1-dioxide	-299.20	-263.93
1,2-dibromocyclooctane	-118.70	-108.05
methylcyclopentane	-106.20	-104.10
cis-1,4-dimethylcyclohexane	-176.60	-185.10
hexachlorobenzene	-44.70	-47.82
methylbenzene	50.40	50.17
cyclobutylamine	41.20	43.57
2-methyl-1,3-dioxane	-397.80	-390.05

2-methyl-3-hydroxypyridine	-84.40	-75.86
dimethylfurazan	107.30	108.24
1,3-dithiol-2-one	-15.00	-34.44
1,3-cycloheptadiene	94.30	96.73
1,2-dibromocyclohexane	-114.30	-109.95
trans-2,4-dimethyl-1,3-dioxolane	-380.50	-391.48
cyclopropene	277.10	284.13
1,3-dithiolan-2-one	-125.90	-129.43
cis-2,3-dimethylthiirane	11.30	3.70
thymine	-328.70	-337.98
cis-1,2-dimethylcyclohexane	-172.10	-174.52
1,3-dithiolane-2-thione	93.80	105.97
thiirane	82.00	74.69
2-methyl-4-hydroxypyridine	-71.70	-85.54
1,1-dimethylcyclohexane	-180.90	-184.11
tetrahydro-3-methyl-thiophene	-60.50	-64.54
2-chloro-1,4-benzendiol	-314.00	-297.30
1h-pyrrole	108.30	108.32
cis-2,4-dimethyl-1,3-dioxane	-425.30	-433.02
1,2,3-trimethylbenzene	-9.50	-12.04
cis-2,4-dimethyl-1,3-dioxolane	-382.60	-391.48
1,3-dioxol-2-one	-418.60	-398.78
cycloheptanone	-247.50	-236.62
2-nitrofuran	-28.80	-39.40
cyclooctanone	-271.60	-252.42
3,5-dimethylisoxazole	-17.90	-9.60

1,3-difluorobenzene	-309.20	-297.03
2,3,5-trichloro-1,4-benzenediol	-339.40	-349.20
benzenethiol	112.40	108.82
2,2,6,6-tetramethyl-4-piperidone	-273.40	-287.26
dihydro-2h-pyran	-125.10	-116.76
cyclohexanethiol	-96.10	-98.83
2,5-dihydrothiophene-1,1-dioxide	-256.20	-231.06
1,4-dichlorobenzene	22.50	17.26
3,3,5,5-tetramethyl-1-pyrazoline	39.20	30.67
dihydro-3,3,4-trimethyl-2,5-furandione	-614.30	-622.31
1,3-dithiane-2-thione	77.50	101.87
4-methylcyclopentene	14.60	7.80
cyclopentylamine	-54.90	-53.31
5-methylisoxazole	34.10	33.32
1-nitropiperidine	-44.50	-41.96
cyclobutene	156.70	162.53
1,3-dioxepane	-346.60	-346.99
hexahydro-2h-azepin-2-one	-246.20	-241.68
tetrahydrofuran	-184.20	-181.67
4,5-dimethyl-1,3-dioxane	-409.10	-412.59
thiophene	114.90	110.74
2,6-dimethylpyridine	58.70	55.77
4-pyridinol	-40.80	-43.21
trans-1,3-dimethylcyclohexane	-176.50	-185.42
3,4,5,6-tetrahydro-3,3,6,6-tetramethylpyridazine	42.00	36.73
4-methylphenol	-125.40	-123.46

1,4-difluorobenzene	-306.70	-308.12
1-fluoro-4-methylbenzene	-147.50	-144.66
oxirane	-52.60	-54.36
1,3-dioxane	-342.30	-340.44
methylcyclohexane	-154.70	-153.25
tetrahydro-2,2,6,6-tetramethyl-4h-thiopyran-4-one	-291.20	-296.98
(nitromethyl)-benzene	30.70	30.41
uracil	-302.90	-299.54
dihydro-2h-thiopyran-3(4h)-one	-133.00	-165.38
2,5-dimethylphenol	-161.60	-159.99
3-methylpyridine	106.40	104.55
1-ethyl-4-nitrobenzene	7.40	-1.48
trans-1,3-dimethyl-cyclopentane	-133.60	-134.73
cyclopentanol	-242.60	-242.53
1,3-dioxolan-2-one	-508.40	-512.48
1,1-dimethylcyclopentane	-138.20	-138.37
cyclooctane	-124.40	-122.21
1,1-dimethylcyclopropane	-8.20	-10.64
1,3,5-trimethylbenzene	-15.90	-16.38
2-methylpiperidine	-84.40	-85.22
2,3-dimethylphenol	-157.20	-157.67
1,3,5,7,9-pentaoxecane	-779.80	-793.35
2-oxetanone	-282.90	-284.14
1,2-dichlorobenzene	30.20	25.85
trans-1,4-dimethylcyclohexane	-184.50	-185.10
cis-1,3-dimethylcyclo-pentane	-135.90	-134.73

4-methyl-1,3-dioxane	-376.90	-383.40
cycloheptene	-9.20	-7.98
aziridine	126.50	127.47
2,3,4,5,6-pentafluoro-1-methylbenzene	-842.90	-840.21
piperidine	-47.20	-47.29
1,3,6-trioxacyclooctane	-467.10	-467.76
1-methylcyclohexene	-43.30	-40.76
1,3-cyclohexadiene	106.20	108.39
2-methyl-1,3-dioxolane	-352.00	-349.21
2,6-dichloro-1,4-benzenediol	-331.50	-320.45
tetramethyldiazetine	150.20	149.28
cis-1,3-dimethylcyclohexane	-184.60	-185.42
1-methylcyclopentene	-3.80	-1.92
3,4-dihydroxy-3-cyclobutene-1,2-dione	-514.50	-440.28
1,4-benzenediol	-265.30	-259.90
cyclooctene	-27.00	-22.67
cyclohexene	-5.00	-2.78
2,5-furandione	-398.30	-402.67
5-amino-3,4-dimethylisoxazole	5.20	-0.50
methylthiirane	45.80	39.39
1,2-dibromocycloheptane	-105.60	-102.75
chlorocyclohexane	-163.70	-166.10
cyclohexanol	-286.20	-290.79
2-methylthiophene	83.50	77.53
chloropentafluorobenzene	-812.30	-807.13
methylenecyclobutane	121.50	122.55

tropolone	-155.40	-155.00
cyclohexane	-123.40	-120.92
1,2,4-trimethylbenzene	-13.80	-16.65
1-methyl-1h-pyrrole	103.10	101.59
5,5-dimethyl-1,3-dioxane	-420.10	-407.78
2,3,5,6-tetrachloro-2,5-cyclohexadiene-1,4-dione	-185.70	-203.10
phenol	-96.40	-91.98
1,4-dioxane	-315.80	-317.72
cyclopentanethiol	-48.00	-48.99
aniline	87.10	87.62
1,2-dibromocyclopentane	-54.90	-62.51
cycloheptane	-118.10	-115.41
2,2,4,4-tetramethyl-1,3-cyclobutanedione	-307.60	-319.80
4-methylthiazole	111.80	104.80
dihydro-3,3-dimethyl-2,5-furandione	-581.70	-592.32
1-hydroxy-2,2,6,6-tetra-methyl-4-piperidinol	-345.10	-383.90
trans-2,3-dimethylthiirane	3.60	3.70

References

(1) Iovanac, N.; Savoie, B. M. Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *J. Phys. Chem. A* **2019**,

(2) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *CoRR* **2015**, *abs/1509.09292*.

(3) Chollet, F., et al. Keras. `https://keras.io`, 2015.

(4) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; `https://www.tensorflow.org/`, Software available from tensorflow.org.