P2O ^{13}C NMR Shift Neural Network: Training Sources and Performance Assessment

DLNMRP2O (Deep Learning NMR, Predicted to [2] Observed) is a neural network model generated with the SpookyNet[1] learning forcefield framework. The P2O network follows an approach similar to the CASCADE[2,3] models suggested by Paton and coworkers. P2O differs in that final ^{13}C shift corrections are based on conformers and Boltzmann weights from the MLHF[4] workflow, additionally while the P2O training set covers similar structures, it is both unique and somewhat larger than that used in either the CASCADE 2.0 model or the *public* training set from the Mestrelabs GCNN model[5] and benefits from inclusion of the majority of natural products available in the NAPROC-13 NMR database.

P2O integration into the ***Spartan*** ML-NMR workflow: conformers, Boltzmann weights, and initial predicted shifts are generated from the 2026 MLHF-NMR protocol. These incorporate MM,[6] QM,[7,8] and several published ML model steps.[9,10,11,12] MLHF provides fast predicted proton and ^{13}C shifts for flexible organic molecules (^{13}C shifts within ≈2.2 ppm) with quality comparable to that calculated from its quantum-chemical precursor, the 2019 "Hehre protocol."[13] Final conformer shifts are corrected with ***P2O*** and weighted using MLHF energies. The resulting ^{13}C shift performance is demonstrably improved. Preliminary assessment follows:

	MLHF only Avg. ^{13}C RMSE	P2O+MLHF Avg. ^{13}C RMSE	MLHF only Avg. ^{13}C MaxE	P2O+MLHF Avg. ^{13}C MaxE
DELTA50	1.73 ppm*	**0.62 ppm***	2.09 ppm*	**0.75 ppm***
246 MNPr	1.94 ppm	**0.85 ppm**	4.62 ppm	**1.95 ppm**
420^{13}CNP	2.23 ppm	**1.66 ppm**	5.34 ppm	**4.11 ppm**

*Unique carbons only

Correcting MLHF-NMR predicted ^{13}C shifts with the P2O neural network improves underlying performance by more than 1.0 ppm for rigid molecules and more than 0.5 ppm for flexible molecules. Against the gold-standard DELTA50[14] validation set, P2O+MLHF performs at 0.62 average RMSE and 0.75 ppm average maximum error; to our knowledge this is the highest DELTA50 performance reported for predicted NMR. Validated against the 420^{13}CNP (420 diverse, flexible structured from the natural products literature), P2O+MLHF performs at an average ^{13}C RMSE of 1.66 ppm and an average maximum error of 4.11 ppm.

While the improvement in predicted ^{13}C shifts is significant, the target application is accurate diastereomer analysis (complete stereochemical assignment), addressing a central challenge for conformationally flexible natural products. Diastereomer analysis for 100 natural products (ranging from 2–11 chiral centers) with X-ray crystal structure confirmation from the Cambridge Structure Database[15,16] supports the utility of MLHF in stereochemical assignment. MLHF confirms the correct diastereomer with an 82% success rate (confirming or favoring at rate of 92%). P2O+MLHF performance is superior, confirming the correct diastereomer with a 94% success rate (confirming or favoring at a rate of 100%).

In evaluation of identifying the correct diastereomers, the following metrics* were used:

C= **Confirms** with a **DP4** score of ***85% or higher*** and an RMSE of 2.2 ppm or lower
F= **Favors** with a **DP4** score of ***50–85%*** and RMSE of 2.0 ppm or lower
S= **Suggests** with a **DP4** score of ***15–50%*** and an RMSE or 1.8 ppm or lower
R= **Refutes** with a **DP4** score of ***0–15%*** irrespective of RMSE

*Metrics for stereochemistry: DP4, RMSE, additionally reviewed: MAE, and max errors at chiral and adjacent centers.

Tables 1 and 2 provide a summary of MLHF[4] and P2O+MLHF diastereomer analysis assessment. Both approaches successfully confirm the correct diastereomer in the majority of cases, but *only* the P2O+MLHF appears to either confirm or favor the correct diastereomer in *all* natural products reviewed:

Table 1. MLHF-NMR ^{13}C Diastereomer assessment on 100 flexible natural products

Outcome	MLHF	avg. DP4	avg. RMSE	avg. Max Error	avg. MAE
confirmed C (%)	82%	98.5%	1.56 ppm	3.81 ppm	1.21 ppm
favored F (%)	10%	70.8%	1.60 ppm	3.58 ppm	1.29 ppm
suggested S (%)	5%	27.9%	1.61 ppm	3.77 ppm	1.25 ppm
refuted R (%)	3%	8.8%	2.08 ppm	5.06 ppm	1.70 ppm
C (%)+ F (%) =	**92%**				

Table 2. P2O+MLHF-NMR ^{13}C Diastereomer assessment on 100 flexible natural products

Outcome	P2O+MLHF	avg. DP4	avg. RMSE	avg. Max Error	avg. MAE
confirmed C (%)	94%	98.9%	0.78 ppm	1.87 ppm	0.61 ppm
favored F (%)	6%	69.9%	0.72 ppm	1.63 ppm	0.57 ppm
C (%)+ F (%) =	**100%**				

In the more challenging task of stereochemical assignment, the P2O+MLHF workflow either confirms (94%) or favors (6%) the correct complete stereochemistry in all 100 reviewed natural products. All structures included in the full diastereomer analysis were confirmed via X-ray diffraction and are represented in the CSD. The full set of 100 NPs including diastereomers with results from both MLHF and P2O+MLHF workflows ("100DA") as well results from DELTA50, 246MNPr and the 420^{13}CNP benchmarks are available for download/review (linked below).

DLNMRP2O Training. Three main sources comprise the training data* (≈38.5k structures with assigned experimental ^{13}C shifts); ≈45% originates from NAPROC-13,[17,18,19] ≈35% from NMRShiftDB2,[20,21] and ≈10% from the SNPD[4,12] (Spartan Natural Products Database). Additional training data was sourced from JEOL's CH-NMR-DB[22] collection as well as open source articles from the natural products literature. Finally, contributions of both training molecules and structure verification were provided by José L. López-Pérez (University of Salamanca, Spain), Masaru Hashimoto (Hirosaki University, Japan), and Novriyandi Hanif (IPB University, Indonesia).

DELTA50 = 50 diverse rigid organic small molecules (gold standard predicted shift validation set)
246 MNPr = 246 rigid marine natural products (DLNMR1 validation set)
420^{13}CNP = 420 diverse flexible natural products (MLHF-NMR validation set)
100DA = 100 natural products with diastereomer analysis (10 each from 2–11 chiral centers)

* The *current* training source includes ≈42k structures. With duplicates removed and a ^{13}C shift quality filter applied, the actual training count is ≈38.5k. Beyond this 42k group, several thousand *additional* natural products and organic molecules with reasonable quality experimental ^{13}C shifts are available. Intuition suggests that at this point, additional performance gains are more likely accessible from smaller, targeted training on molecules with missing or rare structural moieties (rather than training on more structures from the same general sources). Through collaboration with contacts at KNU and Enamine (Ukraine), we're reviewing novel synthetic organic structures for potential targeted retraining candidates.

References

[1] Unke, O. T.; Chmiela, S.; Gastegger, M.; Schütt, K. T.; Sauceda, H. E.; Müller, KR. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat Commun.* **2021**, *12*, 7273. https://doi.org/10.1038/s41467-021-27504-0

[2] Guan, Y.; Sowndarya, S. V. S.; Gallegos, L. C.; St. John, P. C.; Paton, R. S. Real-Time Prediction of ^{1}H and ^{13}C Chemical Shifts with DFT Accuracy Using a 3D Graph Neural Network. *Chem. Sci.* **2021**, *12*, 12012–12026. https://doi.org/10.1039/D1SC03343C

[3] Bhadauria, A.; Feng, Z.; Popescu, M.; Paton, R. S. CASCADE-2.0: Real-Time Prediction of 13C-NMR Shifts with Sub-ppm Accuracy. *ChemRxiv.* **2025**, https://doi.org/10.26434/chemrxiv-2025-r8m9m

[4] Ohlinger, W.; Klunzinger, P.; Hehre, T.; Driessen, A.; Deppmeier, B.; Johnson, J.; Ohlinger, C.; Hehre, W.; Schnitker, J.; Uchida, N.; Watanabe, Y.; Cohen, R. D.; Bruno, I. J.; Hashimoto, M.; Kanehara, R.; Herzon, S. B.; Riseman, H. J.; López-Pérez, J. L. Practical Machine Learning Strategies. 5. Rapid, Accurate Prediction of 1H and 13C NMR Shifts in Conformationally Flexible Organic Molecules. *ChemRxiv* **2026**, https://doi.org/10.26434/chemrxiv.15001441/v2

[5] Williamson, D.; Ponte, S.; Iglesias, I.; Tonge, N.; Cobas, C.; Kemsley, E.K. Chemical shift prediction in ^{13}C NMR spectroscopy using ensembles of message passing neural networks (MPNNs). *J. Mag. Res.* **2024**, *368*, 107795. https://doi.org/10.1016/j.jmr.2024.107795

[6] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *33*, 490. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6%3C490::AID-JCC1%3E3.0.CO;2-P

[7] Binkley, J. S.; Pople, J. A.; Hehre, W. J. Self-Consistent Molecular Orbital Methods. 21. Small Split-Valence Basis Sets for First-Row Elements. *J. Am Chem. Soc*. **1980**, *102*, 939–947. https://doi.org/10.1021/ja00523a008

[8] Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. Self-consistent molecular orbital methods. XXIII. A polarization-function-supplemented basis set for molecules containing second-row atoms. *J. Am Chem. Soc.* **1982**, *104* (10), 2797–2803. https://doi.org/10.1021/ja00374a017

[9] Hehre, T.; Klunzinger, P. E.; Deppmeier, B.; Ohlinger, W.; Hehre, W. Practical Machine Learning Strategies I. Correcting MMFF Molecular Mechanics Model to More Accurately Provide Conformational Energy Differences in Flexible Organic Molecules. *J. Comp. Chem.* **2025**, *46* (1), 70016. https://doi.org/10.1002/jcc.70016

[10] Klunzinger, P.; Hehre, T.; Deppmeier, B.; Ohlinger, W.; Hehre, W. Practical Machine Learning Strategies. 2. Accurate Prediction of ωB97X-V/6-311+G(2df,2p), ωB97M-V/6-311+G(2df,2p) and ωB97M(2)/6-311+G(2df,2p) Energies From Neural Networks Trained From ωB97X-D/6-31G* Equilibrium Geometries and Energies. *J. Comp. Chem.* **2025**, *46* (13), 70129. https://doi.org/10.1002/jcc.70129

[11] Hehre, T.; Klunzinger, P. E.; Deppmeier, B.; Ohlinger, W.; Hehre, W. Accurate Prediction of ωB97X-D/6-31G* Equilibrium Geometries from a Neural Net Starting from Merck Molecular Force Field (MMFF) Molecular Mechanics Geometries. *J. Chem. Info. Mod.* **2025**, *65* (5), 2314–2321. https://doi.org/10.1021/acs.jcim.4c01898

[12] Hehre, T.; Klunzinger, P. E.; Deppmeier, B. J.; Ohlinger, W. S.; Hehre, W. J. Practical Machine Learning Strategies. 4. Using Neural Networks to Replicate Proton and ^{13}C NMR Chemical Shifts Obtained from ωB97X-D/6-31G* Density Functional Calculations. *J. Org. Chem.* **2025**, *90* (32), 11478–11485. https://pubs.acs.org/doi/10.1021/acs.joc.5c00927

[13] Hehre, W.; Klunzinger, P.; Deppmeier, B.; Driessen, A.; Uchida, N.; Hashimoto, M.; Fukushi, E.; Takata, Y. Efficient Protocol for Accurately Calculating ^{13}C Chemical Shifts of Conformationally Flexible Natural Products: Scope, Assessment, and Limitations. *J. Nat. Prod.* **2019**, *82* (8), 2299–2306. https://doi.org/10.1021/acs.jnatprod.9b00603

[14] Cohen, R. D.; Wood, J. S.; Lam, Y-H.; Buevich, A. V.; Sherer, E. C.; Reibarkh, M.; Williamson, R. T.; Martin, G. E. DELTA50: A Highly Accurate Database of Experimental ^{1}H and ^{13}C NMR Chemical Shifts Applied to DFT Benchmarking. Molecules. **2023**, *28* (6), 2449. https://doi.org/10.3390/molecules28062449

[15] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B* **2016**, *72* (2), 171–179. https://doi.org/10.1107/S2052520616003954

[16] Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. Sect. B* **2002**, 58, 380–388. https://doi.org/10.1107/S0108768102003890

[17] López Pérez, J. L.; Therón, R.; del Olmo, E.; Díaz, D. NAPROC 13: a database for the dereplication of natural product mixtures in bioassay guided protocols. *Bioinform*. **2007**, *23* (23), 3256–3257. https://doi.org/10.1093/bioinformatics/btm516

[18] Guerrero De León, E.; Sánchez-Martínez, H.; Morán-Pinzón, J. A.; Del Olmo Fernández, E.; López-Pérez, J. L. Computational Structural Revision of Elaeophorbate and Other Triterpenoids with the Help of NAPROC-13. A New Strategy for Structural Revision of Natural Products*. J. Nat. Prod.* **2023**, *86* (4), 897–908. https://doi.org/10.1021/acs.jnatprod.2c01135

[19] Sánchez-Martinez, H. A.; Morán-Pinzón, J. A.; del Olmo, E.; Adserias Vistué, J. F.; De León, E. G.; López-Pérez, J. L. Structure Revision of Pyranoxanthones via DFT-Assisted ^{13}C NMR Analysis and NAPROC-13 Platform: Diagnostic Markers and Discovery of Unreported Natural Products, *J. Org. Chem.* **2025**, *90* (49), 17240–17251. https://doi.org/10.1021/acs.joc.5c01371

[20] Kuhn, S.; Schlörer, N. I. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2—a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn. Reson. Chem.* **2015**, *53* (8), 582–589. https://doi.org/10.1002/mrc.4263

[21] Kuhn, S.; Kolshorn, H.; Steinbeck, C.; Schlörer, N. Twenty years of nmrshiftdb2: A case study of an open database for analytical chemistry. *Magn. Reson. Chem.* **2024***,* 62 (1), 74–83. https://doi.org/10.1002/mrc.5418

[22] Hayamizu, K. JEOL Natural Products NMR-DB (CH-NMR-NP), **2000**–**2014**. https://ch-nmr-np.jeol.co.jp/en/nmrdb/