HSQC Spectra Simulation and Matching for Molecular Identification

Martin Priessner,[a] Richard J. Lewis,[b] Magnus J. Johansson,[a] Jonathan M. Goodman,[c] Jon Paul Janet,[d] Anna Tomberg[a]**

[a] Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden), email: martin.priessner@astrazeneca.com, anna.tomberg@astrazeneca.com

[b] Department of Medicinal Chemistry, Research & Early Development, Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden)

[c] Centre for Molecular Informatics, The Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, (UK)

[d] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden)

ABSTRACT

In the pursuit of improved compound identification and database search tasks, this study explores Heteronuclear Single Quantum Coherence (HSQC) spectra simulation and matching methodologies. HSQC spectra serve as unique molecular fingerprints, enabling a valuable balance of data collection time and information richness. We conducted a comprehensive evaluation of four HSQC simulation techniques: ACD-Labs (ACD), MestReNova (MNova), Gaussian NMR calculations (DFT), and a graph-based neural network (ML). For with the latter two techniques, we developed a reconstruction logic to combine proton and carbon 1D spectra into HSQC spectra. The methodology involved the implementation of three peak-matching strategies (Minimum-Sum, Euclidean-Distance, and Hungarian-Distance) combined with three padding strategies (zero-padding, peak-truncated, and nearest-neighbor double assignment). We found that coupling these strategies with a robust simulation technique facilitates the accurate identification of correct molecules from similar analogues (regio- and stereoisomers) and allows for fast and accurate large database searches. Furthermore, we demonstrated the efficacy of the best-performing methodology by rectifying the structures of a set of previously misidentified molecules. This research indicates that effective HSQC spectra simulation and matching methodologies significantly facilitate molecular structure elucidation. Furthermore, we offer a Google Colab notebook for researchers to use our methods on their own data (https://github.com/AstraZeneca/hsqc_structure_elucidation.git).

INTRODUCTION

Understanding and characterizing molecular structures is a fundamental task in chemistry, with applications spanning across drug design, synthetic chemistry, and material science. Characterizing complex organic compounds, particularly distinguishing regio- and stereoisomers, remains challenging and often demands a blend of sophisticated analytical techniques. One powerful tool to navigate this complexity is nuclear magnetic resonance (NMR) spectroscopy. Structure elucidation often requires data from several sources. For example, combining a 1D NMR technique such as a proton (1H) or carbon (13C) spectrum and 2D methods like Heteronuclear Single Quantum Coherence (HSQC) or/and Heteronuclear Multiple Bond Correlation (HMBC) can be beneficial.[1] However, practical limitations such as overlapping peaks, impurities, or product mixtures can introduce sources of error rendering the elucidation of the correct structure a highly challenging task. This is reflected in a growing number of natural and synthetic products with incorrectly assigned structures.[2–9]

HSQC is a 2D NMR technique based on linking proton (^{1}H) and carbon (^{13}C) chemical shifts, offering a valuable balance between information content and data collection time. These spectra are contour plots that reveal high-intensity spots or "peaks", denoting direct C-H correlations within the target molecule, serving as a unique molecular fingerprint. Although information rich, the sparse, pseudo-discrete 2D nature of these fingerprints make straightforward comparisons between HSQC spectra challenging due to issues with point matching.[10] The ability to evaluate spectral similarity, that is how well two spectra match, is a basic requirement for comparison of newly-obtained spectra with existing data. While for 1D techniques, lots of work has been done in this area,[11–21] literature on effective comparisons of 2D spectra is largely absent. In this study, we address this deficiency by focusing on HSQC data, aiming to investigate, develop, and evaluate

diverse methodologies for 2D spectral matching, with the ultimate objective of enhancing accuracy and efficiency in tasks such as compound identification and database searches.

There have been some attempts to tackle these challenges. For instance, Cottrell et al. introduced SMART, a technique that leverages a neural network trained on HSQC spectra, to enhance natural products deduplication by identifying clusters of similar compounds.[22,23] DeepSAT developed from the same group employs a neural network-based scaffold prediction system based on chemical features from HSQC spectra.[24] Another 2D method, developed by Sarotti et al., is based on an artificial neural network (ANN-PRA)[25] trained on pattern recognition descriptors of DFT-calculated and experimental HSQC data, which could correctly identify mischaracterized structures from the literature. These approaches, though novel and effective to a certain extent, do bring their own limitations, especially as the performance of a neural network can depend heavily on the diversity of structures included in the training set. As such, they may not fully address the complexities associated with NMR data, such as instrument variability, experimental conditions, and the symmetry or chirality effects of molecules.

Alternatively, a more general statistical approach relying on peak matching is required. A few approaches can be found in the literature that match and compare HSQC spectra. Porzel et al. and Pretsch et al. both used a grid-based approach that divides the spectrum into a grid of a given resolution and assigns it with zeros and ones based on the absence/presence of a peak, resulting in a binary fingerprint for a given molecule.[26,27] This algorithm provides a quick way for deduplicating spectra from databases. However, this approach's lack of peak-to-peak matching and sensitivity to grid resolution can result in decreased accuracy when compared to a one-to-one peak matching technique. Pierre et al. later developed a discrete self-adaptive differential evolution (dSADE)[28] approach and improved the runtime speed of one-to-one peak matching methodology

with an affinity matrix approach.[29] Nonetheless, even this improved algorithm takes roughly 1 hour to match two spectra with 100 peaks (on a 1.8 GHz dual core CPU) making it unsuitable for large database searches.

A common complication for these algorithms stems from the variance in peak numbers across different spectra.[28] Even for the same molecule, the number of peaks can vary due to experimental setup or spectral simulation technique used. For example, in experimental data, some peaks can disappear among the noise, or the measured spectrum can contain numerous exceptions to rule-based splitting patterns, which are dependent on the specific molecular structure and the resolution of the NMR instrument. In simulated spectra, all depends on the type of method used. Most common ways to simulate HSQC data are (1) commercial software like ACD Labs (ACD) or MestReNova (MNova), (2) quantum chemical calculations with for example Density Functional Theory (DFT), and (3) machine learning (ML) models.[30–32] ACD and MNova create their 2D spectra using proprietary algorithms, which include an ensemble of techniques such as calculating chemical shifts based on databases of empirical data, implementation of Hierarchical Organization of Spherical Environments (HOSE) and neural network predictions. Finally, the 1D generated outputs of these software solutions reconstruct a simulated 2D spectrum in the form of a set of peak coordinates in the ^{1}H-^{13}C space, representing the molecular fingerprint of the compound. DFT calculations compute proton (^{1}H) and carbon (^{13}C) chemical shifts for a given molecular structure. A ML model such as the one implemented in this research is based on a graph neural network trained on the NMRShiftDB2 database to reproduce experimental ^{1}H and ^{13}C chemical shifts of small druglike molecules. While commercial solutions provide algorithms that reconstruct the 2D spectrum from computed ^{13}C and ^{1}H NMR chemical shifts, DFT and ML leave this part up to the user. Considering the multitude of sources to obtain HSQC spectra, it is no surprise that the

number of spectral peaks will vary. Hence, this necessitates padding strategies to accommodate this mismatch in peak numbers. Evaluating the most efficient strategy becomes paramount, particularly when it comes to tasks like database searching or correctly identifying molecular structures.

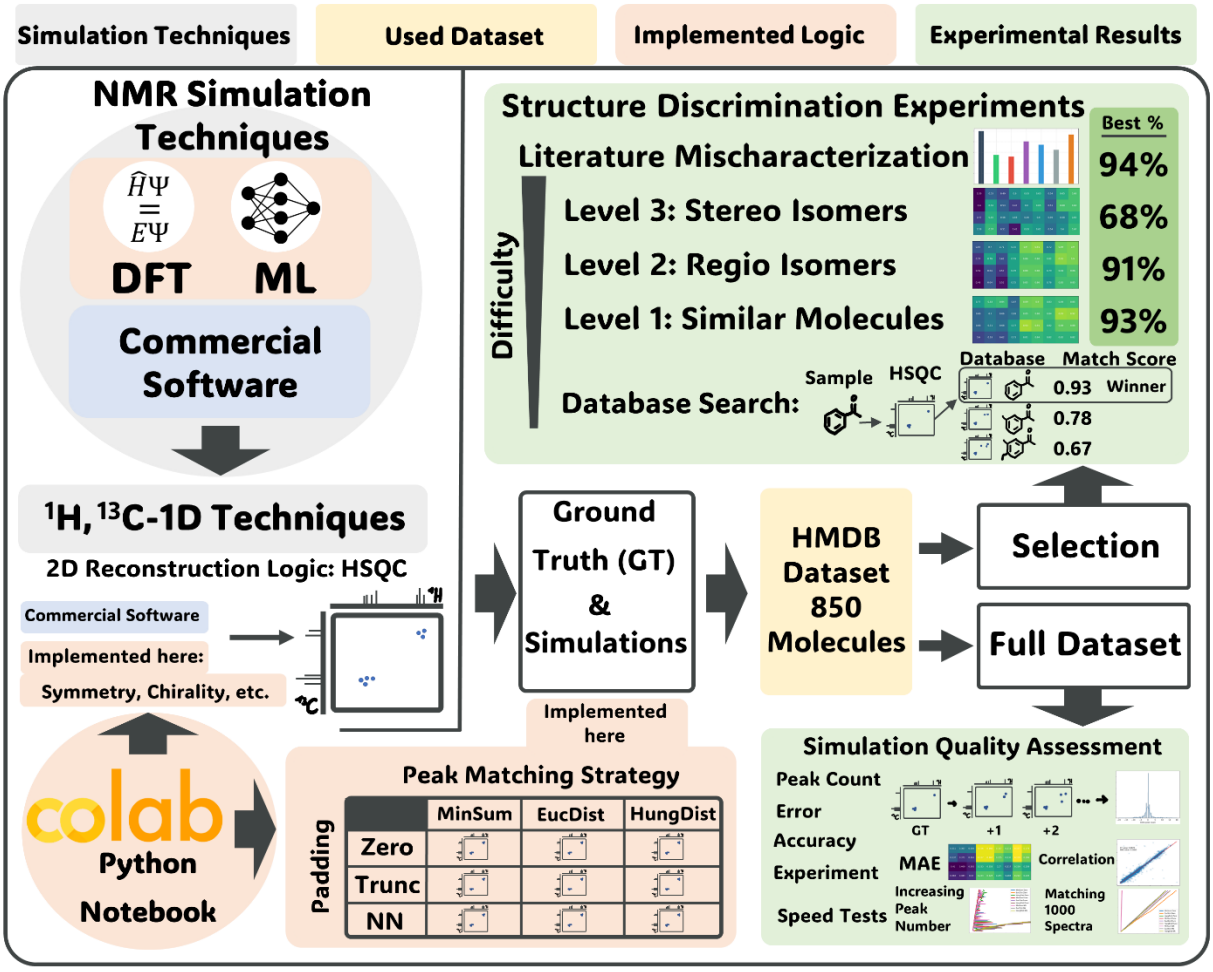


Figure 1. Schematic overview and results of the developed methodology for structure elucidation using HSQC NMR simulation techniques.

In this study, we have developed and rigorously evaluated four spectral simulation techniques, paired with three padding and three matching strategies, benchmarking them across varied tasks and quality metrics. An overview of our work, contributions, and findings is presented in **Figure**

1. The upper-left quadrant outlines the four simulation techniques employed: two commercially available solutions (ACD, MNova), Gaussian[33] NMR calculations (DFT), and a machine learning (ML) method.[30] Additionally, we have implemented an algorithm to reconstruct HSQC spectra from 1D NMR shifts for DFT and ML, taking into account critical factors like symmetry and chirality. Complementing this, we designed a set of strategies that incorporate three peak-matching and three peak-padding techniques to address peak mismatches. Our results provide various quality assessment metrics, such as peak count errors, shift accuracy, and computation speed for the matching/padding techniques. The methodology's versatility was successfully confirmed through evaluations involving database search and structure discrimination tasks, with challenges scaling from similar molecules (Level 1) to regioisomers (Level 2) and finally, to stereoisomers (Level 3). The practical utility of our tool was shown through successful application to previously misassigned molecules from literature as curated from Sarotti et al.[25]. We also provide a Google Colab notebook which houses these tools for convenient matching of simulated and experimental data, thus facilitating structure elucidation. The notebook can be accessed via [https://github.com/AstraZeneca/hsqc_structure_elucidation.git].

RESULTS AND DISCUSSION

We consider a number of different tasks intended to cover practical applications of HSQC in structure elucidation in order to benchmark different simulation and spectral similarity methods:

1. First, we assess how faithfully various simulation techniques can reproduce experimental HSQC spectra.

2. Second, we assess how effectively different spectral similarity methods are for database retrieval tasks with both real and simulated spectra.

3. Next, we assess how well these methods work for structure disambiguation by testing if they distinguish between increasingly-similar molecules: similar size decoys, regiosiosmers and finally stereoisomers.

4. Finally, we apply these methods to previous-identified instances of structural mischaracterization from the literature.

In this study, we evaluated four distinct HSQC spectrum simulation techniques: two commercial software options (ACD and MNova), DFT calculations using Gaussian16, and a ML method (SGNN[30]). In contrast to the two commercial solutions, the shifts of the DFT and ML network require a C-H correlation step to generate a 2D HSQC spectrum, a process that requires an additional layer of computational interpretation. This process must respect molecule symmetry and chirality to perform peak splitting or peak consolidation. In an HSQC NMR experiment, the number of peaks appearing in the spectrum correspond to the unique carbon-hydrogen pairs present in the molecule. Consequently, in symmetric molecules, peaks of distinct H-C couplings can coincide in the same position. However, in chiral compounds, these symmetric peaks may experience different chemical environments and, as a result, not merge into a single peak. We developed an approach for this mapping that takes into account the distance between symmetric atoms and the chiral center, with some additional logic implemented to make the chirality split reliant on the difference in proximity to surrounding functional groups and the chiral center. A comprehensive explanation of the implemented HSQC generation logic is provided in the **Methodology** section.

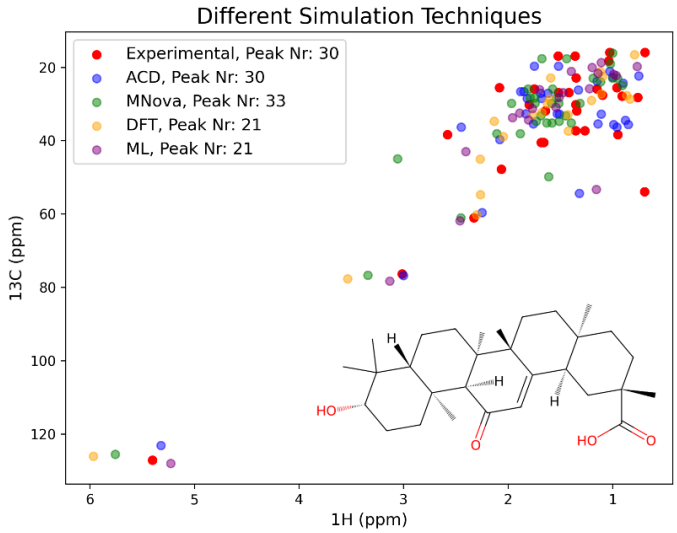


Figure 2. Spectra comparison of experimental data (red) compared to the four different simulation techniques for the metabolite Glycyrrhetinic acid.

Notably, the HSQC reconstruction logic implemented here and the logic from the commercial software solution result in different outcomes. This divergence is due to the different strategies each approach uses in handling molecule symmetry and chirality, among other factors, which eventually lead to a difference in the number of peaks generated (see illustrative example of Glycyrrhetinic acid in **Figure 2**). To initiate our methodological investigation, we started with an examination of the quality variations among the different simulation techniques applied to the public HMDB database,[34] containing 848 molecules. Our evaluation criteria focused on discrepancies in peak count and the accuracy of shift prediction, as described in the forthcoming sections.

Reproduction Accuracy

Spectral Peak Number Comparison

To get a better understanding of the performance of different simulation techniques (ACD, MNova, DFT, ML), we tested their ability to generate the same number of spectral peaks as experimental data from HMDB.

This experiment involved generating spectra for the chosen dataset and comparing the simulated and experimental (ground-truth (GT)) number of peaks (**Figure 3**).

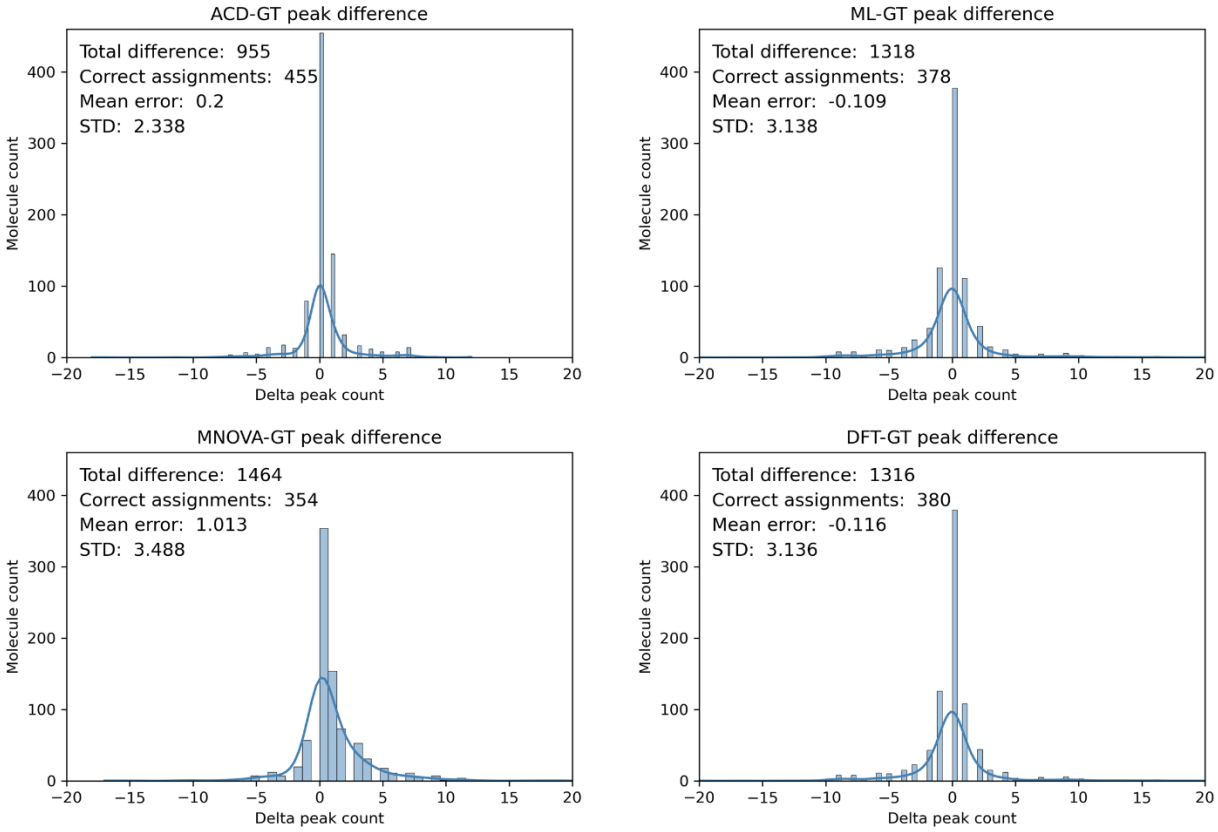


Figure 3. Peak number mismatch between the experimental data and various simulation methods (ACD, MNova, DFT, ML) plotted as histograms with bins indicating the peak number differences between the ground-truth (GT) and simulated spectra. The total difference, correct assignments, mean error, and standard deviation are provided as inset text within each subplot.

It is important to note that the accuracy of a simulation doesn't directly correspond to the precise count of peaks. In the domain of spectral reconstruction, the aim is often to suggest all possible peaks, not all of which may be present or detectable in an experimental setting. In fact, low

intensity peaks might be lost in the noise of experimental data, making the simulated and experimental data appear dissimilar when they are, in fact, both accurate. Therefore, all techniques made mistakes for certain molecules. This is because the GT spectra contain numerous exceptions to rule-based splitting patterns, which are dependent on the specific molecular structure and the resolution of the NMR instrument.

Among the four techniques, ACD displayed the highest accuracy, correctly identifying the number of peaks for 455 out of 848 molecules, while MNova introduced the greatest number of errors in this task (correct/total: 354/848). Meanwhile, spectra reconstructed from DFT calculations (379/848) and ML predictions (378/848), demonstrated a performance level between the two commercial software solutions. Interestingly, the commercial software was more prone to overestimating the number of peaks compared to the experiment. Specifically, MNova showed the highest mean error, followed by ACD. Overall, our results indicate that the peak reconstruction performance of each technique is nearly equivalent when applied to this dataset. The performance metrics were assessed across varying 'delta windows', which are ranges of acceptable deviation from the correct number of peaks. For a more granular breakdown of performance across different delta windows, see **Supplementary Figure S1**. Furthermore, six representative examples of the failure mechanisms (e.g., symmetry/chirality misinterpretation) of the four techniques are illustrated in **Supplementary Figure S2** as well as one detailed breakdown of error analysis of one chiral compound in **Supplementary Figure S3** that results in producing different numbers of peaks for the different simulation techniques. We hypothesized that a varying number of peaks in the simulated spectra could lead to misinterpretation of the underlying molecular structure. Additionally, we anticipated that discrepancies in the number of peaks between two spectra could result in systematic performance shifts when comparing different simulation techniques.

Consequently, we continued our investigation by developing strategies to address mismatches in peak numbers between spectra.

HSQC Simulation Shift Error Evaluation

As demonstrated in the previous experiment, the number of measured experimental GT peaks can often differ from the simulated spectra. In order to compare various simulation techniques, we have incorporated three peak matching methodologies along with three different padding approaches.

Matching methodologies define the principles by which each peak of one spectrum finds its "match" on another spectrum. The three approaches examined here are minimum distance (MinSum), Euclidian distance (EucDist) and Hungarian distance (HungDist). MinSum organizes the peaks of the reference spectrum by their sum of normalized H and C shifts and aligns them with the sorted peaks from the comparison spectrum. EucDist utilizes a nearest neighbor search approach. Finally, HungDist conducts peak matching through the Hungarian algorithm, a combinatorial optimization technique that addresses the assignment problem (Kuhn-Munkres algorithm[35]).

Handling discrepancies in peak numbers can be accomplished in a variety of ways, we have implemented three: zero-padding (Zero), peak truncation (Trunc), nearest-neighbor double-assignment (NN). Zero-padding means that a shorter peak list is extended with a fictional peak at (0,0) that all the unmatched peaks are assigned to. The peak truncation method focuses on optimally matching peaks in the shorter peak list of a spectrum to another, truncating and ignoring any remaining peaks in the longer peak list. Lastly, NN indicates that unmatched peaks from one peak list are paired with their nearest neighbor via double assignment to the closest

available peak. For additional information on these algorithms including a visual illustration of the padding strategies, see **Methodology Section** and **Supplementary Figure S4**.

In short, these matching algorithms and padding techniques create nine combinations (MinSum-Zero, EucDist-Zero, HungDist-Zero, MinSum-Trunc, EucDist-Trunc, HungDist-Trunc, MinSum-NN, EucDist-NN, and HungDist-NN). Next, we investigated how these combinations affected the performance of spectral matching. To achieve this, we matched the simulated HSQC spectra for the HMDB compounds to their GT and computed the normalized mean absolute error (MAE) corresponding to each matching/padding combination (**Figure 4a**). The MinSum-Zero strategy performed worst for all simulation techniques, whereas MNova with HungDist-Trunc or EucDist-Trunc gave the best results overall, although the differences between the top candidates were small. ACD was the second-best simulation technique followed by ML and DFT. Furthermore, zero-padding introduced the biggest systematic error for all matching methodologies.

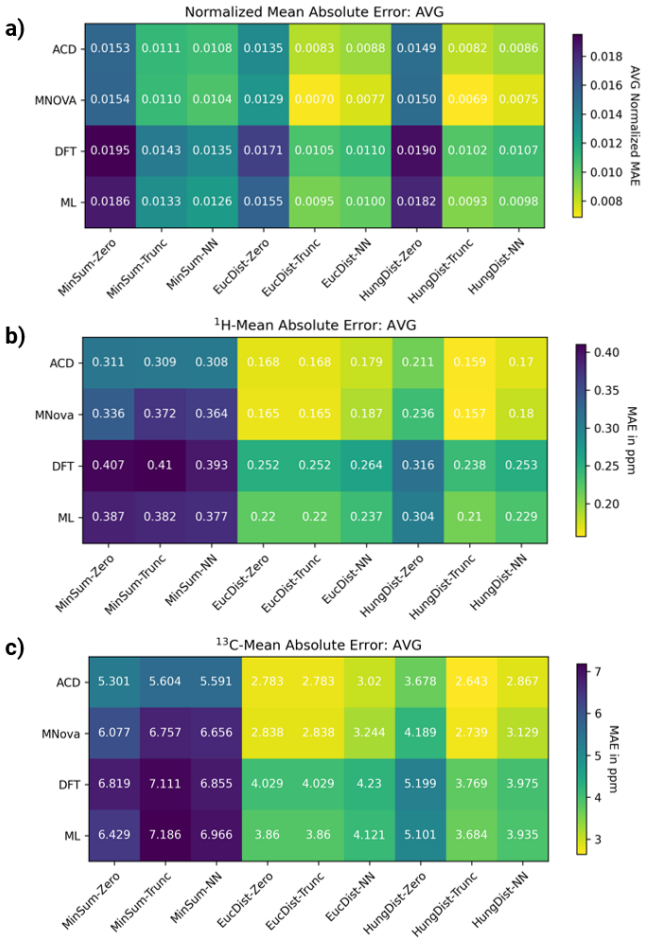


Figure 4. Evaluation of simulation techniques using mean absolute error (MAE) in conjunction with matching algorithms. (a) Comparison of normalized MAE, (b) heatmap of average MAE for matched ^{1}H shifts, and (c) heatmap of average MAE for matched ^{13}C shifts.

In a similar fashion, we analyzed the matched ^{1}H and ^{13}C shifts errors individually for each simulation and matching/padding technique. **Figure 4b** and **4c** illustrate the average MAE in a heatmap. For the shift error calculations of spectra with peak mismatches compensated by zero-padding, the points matched with zeros were excluded from the analysis to avoid artificially penalizing the matching points.

The direct correlation plots followed the same trend as the MAE results showing higher R^2 values for methods with low MAE. **Supplementary Figure S5** displays one matching technique (EucDist-NN) for each simulation method.

The results from the two shift calculations reveal that ACD generated the most accurate spectra simulations, with MNova coming in second, exhibiting the smallest errors of approximately 0.157 ppm and 0.159 ppm for ^{1}H shifts and 2.643 ppm and 2.739 ppm for ^{13}C NMR shifts, respectively, when using the HungDist-Trunc method. MinSum performed worst with all padding strategies and all simulation techniques (see **Figure 4b/c**).

Surprisingly, DFT and ML displayed quite similar performance. This could be explained by the fact that the ML network was trained on shifts of real NMR data, which could have led to the better performance observed. Moreover, the DFT calculations for the entire HMDB dataset were considering the ten most populated conformers. Therefore, results could be optimized by increasing the conformational sampling and potentially by an alternate selection of functional and basis set. Overall, the two commercial simulation techniques delivered the highest accuracy across different matching and padding strategies showing the best performance with HungDist-Trunc and EucDist-Trunc.

While predicting NMR shifts as close to experiment as possible is important, it is not the ultimate goal. We hypothesized that it is likely that a relatively high error could be tolerated in various downstream analysis tasks, such as database search and structure identification, so long as the errors were systematic.

Database Search

GT and Noised-GT Spectral Search

While the accurate computational simulation of NMR is a desirable goal, our primary focus is on the characterization of molecular structures. It is essential to acknowledge that the two are not identical – more accurate simulations do not automatically ensure more precise identification, and the converse also holds true.

With this context in mind, we aimed to evaluate the effectiveness of various matching strategies in the realm of database (DB) search, a crucial real-world application for molecule identification. To assess the nine matching/padding techniques, we analyzed their ability to correctly associate the spectrum corresponding to each molecule within the HMDB dataset. The task was to identify the same experimental spectrum to the one from the database. Since this is a relatively unchallenging task, we obtained a 100% retrieval rate, successfully determining the correct spectra for each padding strategy and each matching technique. To make this scenario more realistic and account for spectral changes due to different concentrations or incorrect referencing, we also looked at progressively noised GT spectra. We systematically or randomly varied the shifts of each peak in the spectra to determine the breaking point of the matching algorithms. The method of spectral noising is described in greater detail in the **Methodology Section**. As expected, with increased level of noise, the quality of the matches deteriorated (for random noise see **Figure 5** and for systematic noise see **Supplementary Figure S6**).

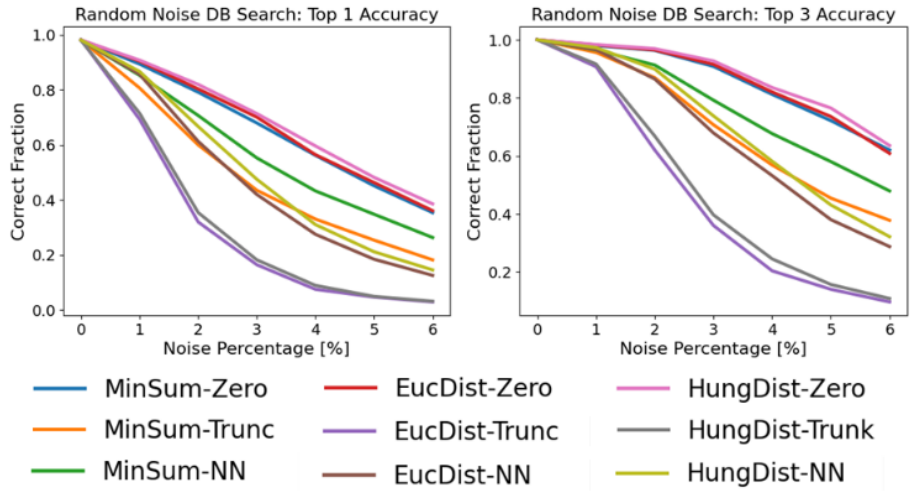


Figure 5. Database search performance with increasing data augmentation. Top-1 and top-3 accuracy of database search with increasing random noised shift of GT spectra.

Systematic noise led to a faster decline in retrieval compared to randomly augmented spectra. Padding strategy had the biggest influence for the database search performance, with zero-padding consistently performing best across the different matching techniques. Peak-truncation was the worst padding strategy. From the matching algorithms, the MinSum method was the most reliable for systematic shifts whereas HungDist performed best for random noise.

Additionally, we benchmarked the speed of the matching algorithms on a spectra database search, aiming to identify efficient and scalable options. The MinSum algorithm was the fastest, while HungDist-Zero showed similar speed with superior matching results and EucDist was consistently slow due to point-to-point matching calculations (for more details see **Supplementary Text 1**).

This benchmarking analysis of the matching algorithms highlights the trade-offs between speed and accuracy for each algorithm. In scenarios where time efficiency is a priority, the MinSum

algorithm offers the fastest performance. However, if the focus is on achieving better matching results without a significant compromise in speed, the HungDist-Zero method is an optimal choice.

GT and Simulated Spectral Search

Next, we assessed the database search on the selected dataset using spectra generated with the four simulation techniques. This serves as a comprehensive comparison of the performance of these techniques to further evaluate their strengths and weaknesses. This task was significantly more challenging due to the errors in simulated peak shifts and peak numbers compared to GT. The best-performing algorithm was HungDist-NN, achieving a top-1 and top-3 accuracy of 56% and 77% for ACD, respectively. Among the simulation techniques, ACD performed the best, followed by MNova and ML. DFT performed notably worse on this task. For the EucDist matching, MNova performed best with Trunc and NN as padding strategy. In contrast to the database search results with real noised spectra, the zero-padded matching strategies were less reliable compared to the nearest-neighbor double-assignment and peak truncation strategies, and MinSum exhibited the weakest performance as a matching technique. The comprehensive results of the top-1 and top-3 benchmark are displayed in **Figure 6**. In the context of this task, the simulated spectra derived from the commercial solutions demonstrated superior performance when applied in conjunction with the HungDist-NN method.

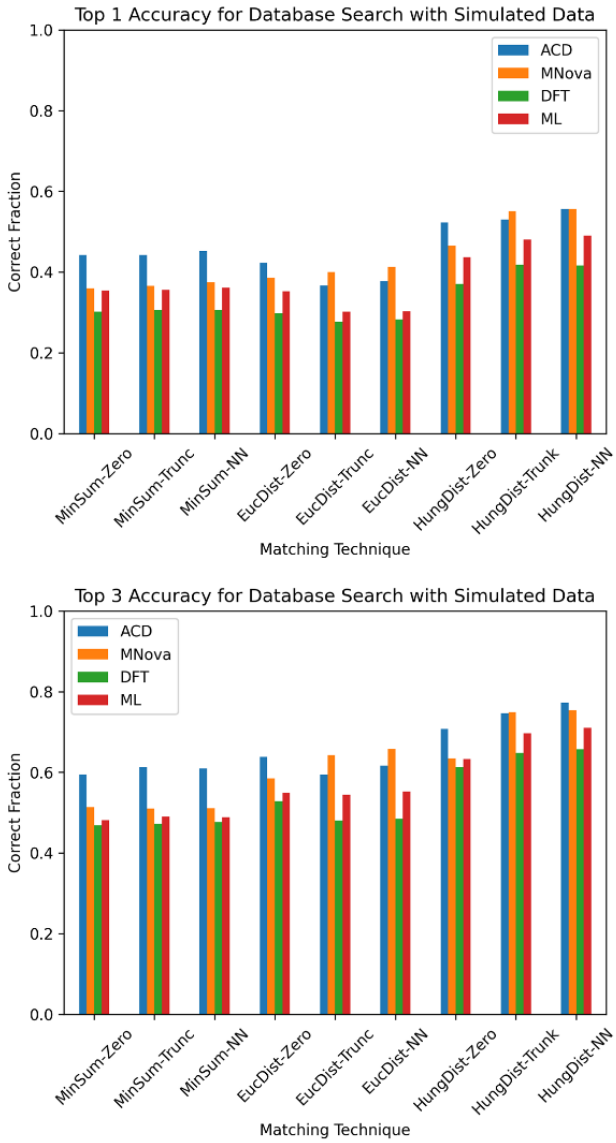


Figure 6. GT database search performance of simulated spectra. Top-1 (top) and top-3 (bottom) retrieval accuracy for the different simulation techniques (ACD, MNova, DFT, ML) of real spectra from the database.

Structure Discrimination

After evaluating the best performing simulation techniques and benchmarking the different peak matching strategies on the task of database search, we gradually increased the complexity of the task. We examined the matching algorithms' ability to identify the correct compound from a set of

similar generated molecules by assessing their spectra similarity to the experimental ground truth (GT) spectrum. Our underlying hypothesis was that the simulated spectrum of the correct molecule would show greater similarity to the GT spectrum compared to the simulated spectra of a decoy. This mimics the real-world use-case of structure disambiguation between similar reaction products. We increased the complexity of the task by creating decoys that were progressively more similar to the original:

- Level 1: Near neighbor augmentation of similar molecular weight
- Level 2: Regioisomer augmentation
- Level 3: Stereoisomer augmentation

We also conducted a control experiment where we systematically assigned the "correct" label to each decoy, making it the reference for comparison with all the others. This approach resulted in a consistent 50% hit rate, which corresponds to random guessing. We displayed the accuracy percentage of each simulation technique with each matching methodology in the form of 2D heatmaps in the following subsections. A visual representation of this methodology is provided in **Figure 7**.

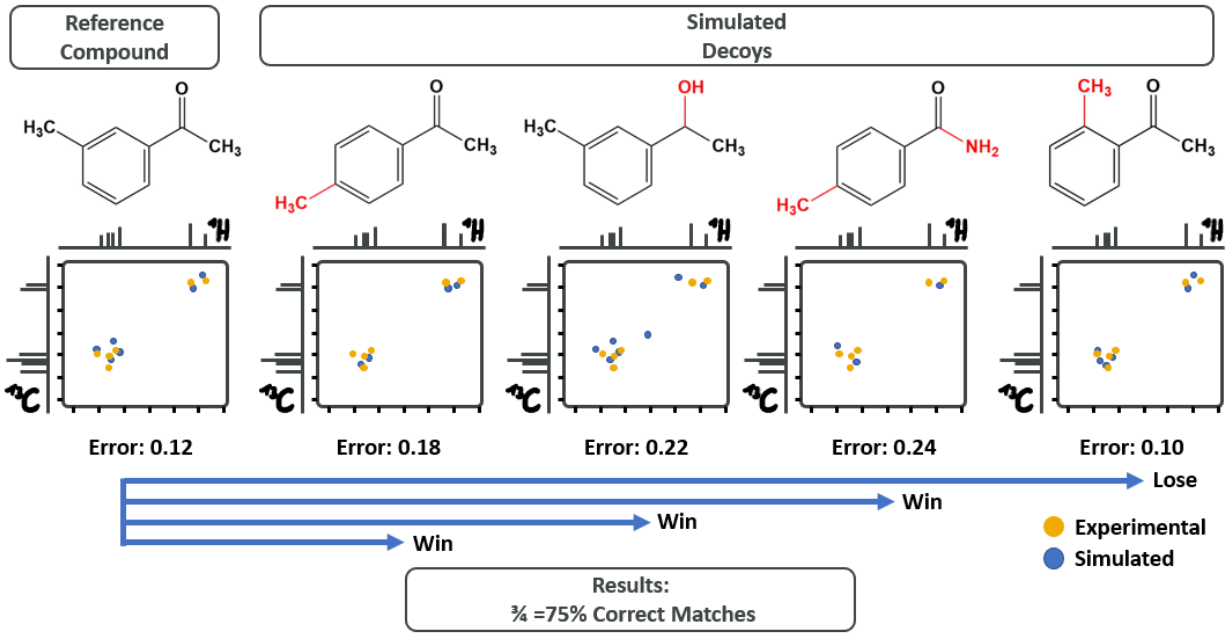


Figure 7. Visual representation of the matching logic for determining the accuracy of correct assignment for the correct molecule given the other analogues.

Level 1: Nearest Neighbor Augmentation

For the first task of distinguishing similar molecules based on simulated versus GT spectra, we selected a representative subset of molecules from the HMDB database. We chose one of the largest molecules from the 17 largest clusters using the Butina method.[36] For each selected molecule, we generated 10 analogues using a near neighbors' molecule generator transformer network,[37] (see the **Methodology Section**).

For the resulting 181 compounds (the transformer model failed to generate all 10 analogues for 3 compounds, resulting in 6 fewer molecules), we calculated each of the four simulated spectra (ACD, MNova, DFT, ML). Each set of molecules was compared to its GT HSQC spectrum by each of the 9 peak-matching/padding techniques resulting in 164 comparison pairs. A heatmap

illustrating the percentages of correctly identified molecules in the 1:1 matching result is presented in **Figure 8.** The control results, displaying a consistent probability centered around 0.5, can be found in **Supplementary Figure S8a**.

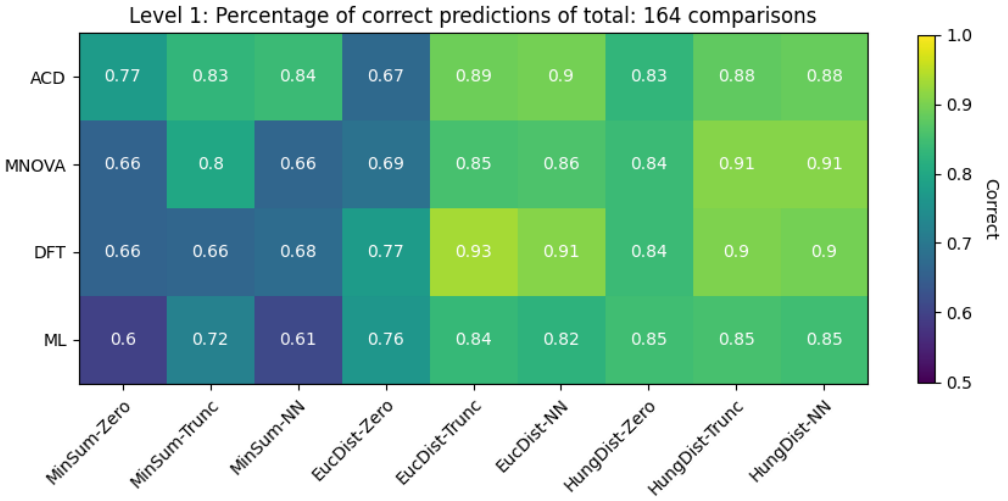


Figure 8: Heatmap results of Level 1: Augmented Molecules. Heatmap illustrating the percentage of correctly identified molecules for the first task of distinguishing similar molecules based on simulated versus GT spectra comparisons. A representative subset of molecules from the HMDB database was used, and various matching and padding strategies were employed. Significant improvements over the baseline random guessing were observed, with the most notable results achieved using HungDist-Trunc/NN matching with MNova and EucDist-Trunc/NN matching with DFT simulated spectra.

Compared to the baseline, these results showed significant improvements depending on the type of matching and padding strategy used. In general, it can be concluded that matching methods and padding strategies are more important than the simulation techniques which (except from MinSum matching) are consistently performing well (>80%) across all padding methods. Moreover, zero-padding showed consistently the weakest performance across the different matching techniques. Highest accuracy was observed for the EucDist-Trunc matching with DFT

(93% accurate) as well as HungDist-Trunc/NN matching with ACD and MNova (91% accurate). ACD results were most robust, displaying over 70% accuracy with all matching algorithms and padding strategies. Furthermore, it can be concluded that zero-padding performs worse compared to peak-truncation or nearest-neighbor double assignment and that ML simulated results performed slightly worse across all matching and padding strategies but still achieved a decent performance.

Level 2: Regioisomer Augmentation

We generated regioisomers by selecting 16 molecules from the largest clusters and manually relocating and reconnecting different side chains and substructures within the molecules to alternative connection points, while maintaining a constant molecular weight. We created five comparison analogues for each molecule. The same logic as in the previous experiment was applied (**Figure 7**), resulting in a total of 80 1:1 comparisons between correct and incorrect molecules. Additionally, we performed a control experiment comparing every molecule with each other within the same category as done in Level 1. **Figure 9** presents a heatmap illustrating the percentages of correctly identified molecules in the 1:1 matching results, while **Supplementary Figure S8b** shows the control results, displaying a consistent probability centered around 0.5.

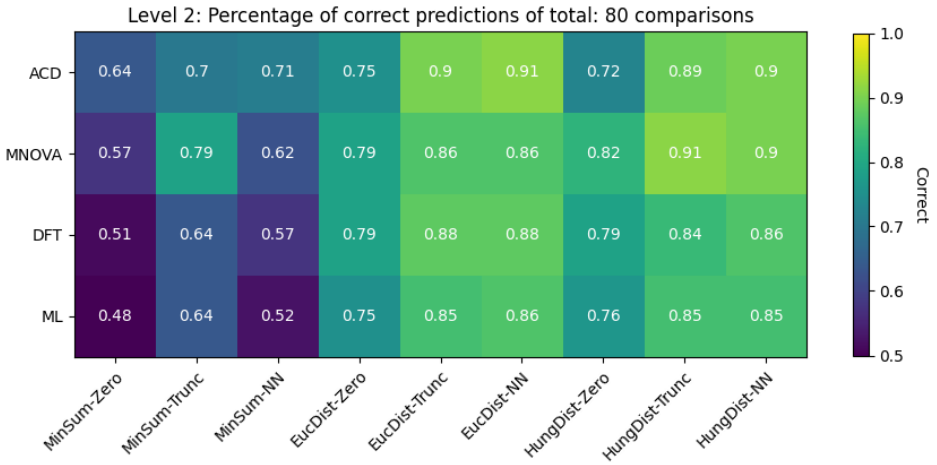


Figure 9. Heatmap results of Level 2: Regioisomer determination. The numbers illustrate the percentage of correctly identified molecules. Significant improvements over the baseline were observed, with the most notable results achieved using HungDist-Trunc/NN matching with MNova and EucDist-Trunc/NN matching with ACD simulated spectra.

In this experiment, comparable accuracy to the level 1 decoys was achieved. This time MNova and ACD (91%) with EucDist-NN and HungDist-Trunc, respectively delivered the best performance, followed by DFT (88%) and ML (86%) using EucDist-NN. Here the same trends were observed as for Level 1. MinSum exhibited the weakest performance for all simulation techniques. In combination with peak truncation as padding strategy it gave the most optimal results. Zero-padding performed consistently weaker compared to the other padding strategies.

Level 3: Stereoisomer Augmentation

For this experiment we selected 14 structures which displayed specific stereo chemistry configurations of which we created between 2 to 6 alternative stereoisomers. Then we followed the same methodology of performing 1:1 comparisons and control experiments. Results are shown in **Figure 10** (and **Supplementary Figure S8c** respectively)**.**

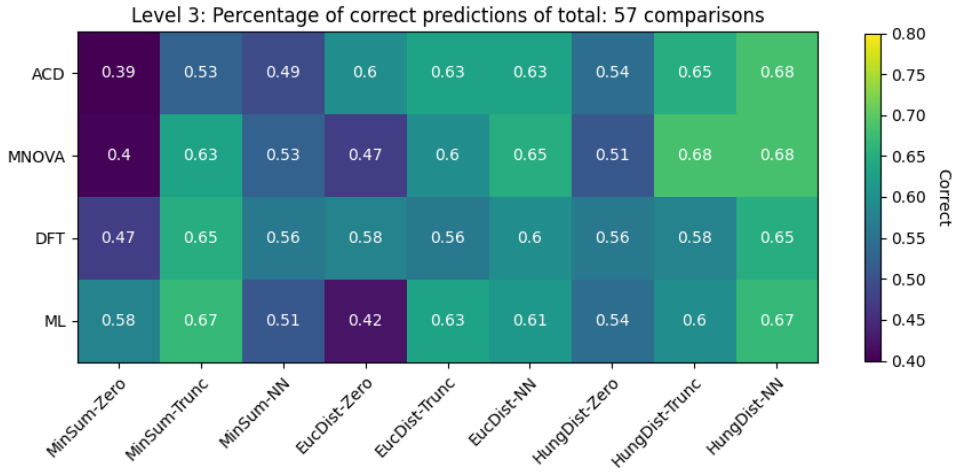


Figure 10. Heatmap results of Level 3: Stereo-isomer determination. The numbers illustrate the percentage of correctly identified molecules. Significant improvements over the baseline were observed, with the most notable results achieved using HungDist-Trunc/NN matching with MNova and EucDist-Trunc/NN matching with ACD simulated spectra. Due to some slightly negative correlation and a reduced sensitivity in this performance level, the color interval range for this illustration was changed from 0.5-1.0 to 0.4-0.8.

In this level, the best performing matching/padding combination across all simulation techniques was HungDist/NN for which all simulation techniques achieved a very similar performance of 65% - 68%. ACD and MNova achieved equivalent top results once again, followed by ML and DFT. MinSum/Trunc performed exceptionally well in this level with ML (67%) as top performer in this category. Zero-padding again displayed the weakest results across all matching techniques.

Correcting Structure Mischaracterizations of Literature Compounds

In a final evaluation, we employed the molecule distinction methodology on real misassigned molecular spectra from literature, curated by Sarotti et al.[25] The tested dataset includes a wide variety of molecules featuring small to medium misassignments of regio- and stereochemistry,

as well as structural errors where the molecules maintain the same molecular weight but exhibit different structures. In addition to the 4 simulation techniques (ACD, MNova, DFT, ML), we also compared the DFT calculations from the original paper with our own DFT results which followed slightly different settings as outlined in the **Methodology Section**. Since our previous results suggested that the HungDist-NN produced the most accurate results across the different levels, we used this matching/padding strategy.

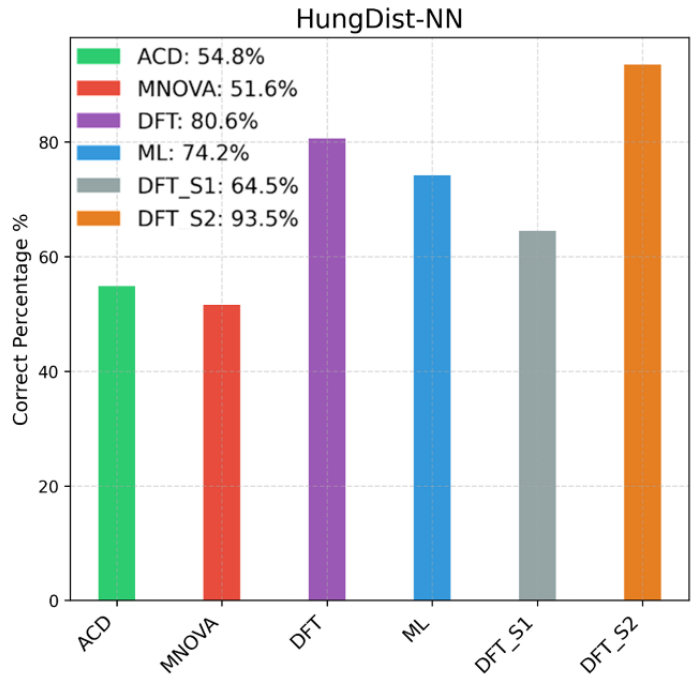


Figure 11. Results of correcting misassigned molecules from the literature. Results of HungDist matching with NN padding with the different simulation techniques (ACD, MNova, DFT, ML, DFT_S1 (Sarotti), DFT_S2 (Sarotti))

In this particular experiment, both DFT and ML methods performed better than the commercial solutions (**Figure 11**). Our DFT results managed to correct 80% of the misassigned molecules and the overall best performing DFT calculations of Sarotti et al. (DFT_S2) attained an accuracy

of 94%. ML demonstrated strong performance as well, achieving 77% accuracy. In contrast, the commercial solutions ACD and MNova performed merely at the level of random chance, with accuracies of 55% and 52% respectively.

The limited performance of ACD and MNova in the final task could potentially stem from their reliance on empirical databases and Hierarchical Organization of Spherical Environments (HOSE) methods. While these approaches are generally reliable for common molecular structures, they may exhibit limitations when encountering novel or complex molecules outside their databases.

Another possible factor could be the deviation of the test data from their training data's distribution, presenting complexities that these tools may not handle optimally. This could explain their relatively lower accuracy in the structure correction task of the Sarotti dataset compared to the experiments performed on the HMDB data. Three illustrative cases of simulation inaccuracies across various algorithms are presented in **Supplementary Figures S9-11**. These examples demonstrate the performance of different simulation techniques when combined with the Hungarian Distance (HungDist) algorithm and Nearest Neighbor (NN) padding, highlighting specific instances of failure in accurately identifying the correct molecule. However, a comprehensive understanding of the performance dynamics of these tools would require further exploration, including an analysis of the source algorithms of the commercial software, which is currently beyond our reach due to access limitations. On the other hand, the consistency of DFT methods across the different tasks is likely due to the intrinsic flexibility of DFT calculations, which are not constrained by predefined databases.

Additionally, the findings highlight the significant role of the appropriate DFT methodology in the performance of simulation techniques. Notably, the two processing methodologies implemented

by Sarotti et al. showcased a substantial performance leap from 65% to 94% between DFT_S1 and DFT_S2. Another remarkable aspect of these results lies in the robust performance of ML, a fact that becomes even more compelling when considering the substantially lower computational costs associated with ML compared to DFT methods.

Conclusion

In this study, we rigorously evaluated the performance and utility of various HSQC simulation techniques and peak-matching strategies in the context of NMR spectral analysis. Our comprehensive assessment covered four key areas:

1. **Reproduction of Experimental HSQC Spectra:** Our analysis revealed that among the four simulation techniques (ACD, MNova, DFT, ML), ACD was the most proficient in accurately reproducing the number of spectral peaks found in experimental HSQC spectra. We further investigated the shift prediction performance of these techniques by incorporating three peak-matching methodologies—MinSum, EucDist, and HungDist—alongside three padding approaches—Zero, Trunc, NN. The commercial solutions, ACD and MNova, produced the most accurate shift prediction on this task.
2. **Effectiveness in Database Retrieval:** Our first practical focus was on the real-world application of molecular identification through database search. We evaluated all nine matching and padding combinations for their ability to correctly identify spectra within the HMDB dataset. While a 100% retrieval rate was achieved in ideal conditions, the performance varied when introducing systematic and random noise into the spectra. Zero-padding emerged as the most robust padding strategy, particularly under conditions of random noise, whereas MinSum was the most reliable matching algorithm for systematic peak shifts.

3. **Structure Disambiguation:** We tested these matching/padding strategies on progressively more challenging tasks, demonstrating their capability to correctly identify molecules among analogues (structural isomers with the same molecular weight or different regio- or stereoisomers). We observed higher sensitivity to the peak matching and padding methods than the choice of simulation techniques overall, emphasizing the importance of this aspect and the novel contributions of this work. Remarkably, maximum performance reached impressive levels of 93%, 91%, and 68% for the three respective tasks, further highlighting the efficacy of our peak matching and padding methodologies.
4. **Correction of Previously Misidentified Structures:** In a practical application of our methodologies, we were able to resolve ambiguous structural assignments in previously mischaracterized molecules. Here, DFT with the HungDist-NN strategy showed the most consistent and robust performance.

Additionally, we provide a Google Colab notebook on our github page which allows to run the ML NMR prediction and includes instructions on generating the simulated spectra with the commercial software (https://github.com/AstraZeneca/hsqc_structure_elucidation.git). The notebook provides instructions on processing real spectra and conducting similarity comparisons using the algorithms implemented here for distinguishing different molecules. This interactive tool is designed to provide hands-on learning and enhance the user's understanding of the methodologies used. Furthermore, the data that support the findings of this study are openly available at (https://doi.org/10.5281/zenodo.8403376).

Ultimately, this research contributes valuable insights into the performance and utility of various simulation techniques and peak-matching strategies. It not only paves the way for improved

efficiency and accuracy in the field of NMR spectral analysis but also broadens its applications in molecular identification and structural elucidation.

ASSOCIATED CONTENT

Supporting_Information.pdf: This file contains all the additional supporting figures as well as methodology, experimental design, additional experiments (Algorithm Speed Test, DFT Benchmarking) and explanation of the HSQC Generation Technique from 1D NMR data.

AUTHOR INFORMATION

Corresponding Authors

Martin Priessner - Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden), email: martin.priessner@astrazeneca.com

Anna Tomberg - Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden), email: anna.tomberg@astrazeneca.com

Authors

Magnus J. Johansson - Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden)

Richard J. Lewis - Department of Medicinal Chemistry, Research & Early Development, Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden)

Jonathan M. Goodman - Centre for Molecular Informatics, Department of Chemistry University of Cambridge, Lensfield Road, Cambridge CB2 1EW, (UK)

Jon Paul Janet - Molecular AI, Discovery Sciences, R&D, AstraZeneca, Pepparedsleden 1, 43183 Mölndal (Sweden)

Author Contributions

Martin Priessner: data curation, investigation, formal analysis, writing – original draft. Jon Paul Janet and Anna Tomberg: conceptualization, funding acquisition, supervision, writing – review and editing. Richard J. Lewis, Magnus J. Johansson and Jonathan M. Goodman: supervision and writing – review and editing. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

We gratefully acknowledge AstraZeneca for their support and funding of the Postdoctoral position, instrumental in the success of this research.

ABBREVIATIONS

ACD, ACD Labs, MNova, MestroNova, ML, Machine Learning, NMR, nuclear magnetic resonance, HSQC, Heteronuclear Single Quantum Coherence, HMBC, Heteronuclear Multiple Bond Correlation, MinSum, minimum distance, EucDist, Euclidian distance, HungDist,

Hungarian distance, Zero, zero-padding, Trunc, peak truncation, NN, nearest-neighbor double-assignment.

REFERENCES

(1) Claridge, T. D. W. *High-Resolution NMR Techniques in Organic Chemistry: Third Edition*; Elsevier Inc., 2016.

(2) Chhetri, B. K.; Lavoie, S.; Sweeney-Jones, A. M.; Kubanek, J. Recent Trends in the Structural Revision of Natural Products. *Nat Prod Rep* **2018**, *35* (6), 514–531. https://doi.org/10.1039/C8NP00011E.

(3) Nicolaou, K. C.; Snyder, S. A. Chasing Molecules That Were Never There: Misassigned Natural Products and the Role of Chemical Synthesis in Modern Structure Elucidation. *Angew. Chem.* **2005**, *44* (7), 1012–1044. https://doi.org/10.1002/ANIE.200460864.

(4) Nicolaou, K. C.; Hale, C. R. H.; Nilewski, C.; Ioannidou, H. A.; Elmarrouni, A.; Nilewski, L. G.; Beabout, K.; Wang, T. T.; Shamoo, Y. Total Synthesis of Viridicatumtoxin B and Analogues Thereof: Strategy Evolution, Structural Revision, and Biological Evaluation. *J. Am. Chem. Soc.* **2014**, *136* (34), 12137–12160. https://doi.org/10.1021/JA506472U.

(5) Xiao, Q.; Young, K.; Zakarian, A. Total Synthesis and Structural Revision of (+)-Muironolide A. *J. Am. Chem. Soc.* **2015**, *137* (18), 5907–5910. https://doi.org/10.1021/JACS.5B03531/SUPPL_FILE/JA5B03531_SI_003.CIF.

(6) Zhu, L.; Liu, Y.; Ma, R.; Tong, R. Total Synthesis and Structural Revision of (+)-Uprolide G Acetate. *Angew. Chem.* **2015**, *54* (2), 627–632. https://doi.org/10.1002/ANIE.201409618.

(7) Nicolaou, K. C.; Shah, A. A.; Korman, H.; Khan, T.; Shi, L.; Worawalai, W.; Theodorakis, E. A. Total Synthesis and Structural Revision of Antibiotic CJ-16,264. *Angew. Chem.* **2015**, *54* (32), 9203–9208. https://doi.org/10.1002/ANIE.201504337.

(8) Maier, M. E. Structural Revisions of Natural Products by Total Synthesis. *Nat Prod Rep* **2009**, *26* (9), 1105–1124. https://doi.org/10.1039/B809658A.

(9) Suyama, T. L.; Gerwick, W. H.; McPhail, K. L. Survey of Marine Natural Product Structure Revisions: A Synergy of Spectroscopy and Chemical Synthesis. *Bioorg Med Chem* **2011**, *19* (22), 6675–6701. https://doi.org/10.1016/J.BMC.2011.06.011.

(10) Huang, J.; You, S. Point Cloud Matching Based on 3D Self-Similarity. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* **2012**, 41–48.

(11) Kutateladze, A. G.; Krenske, E. H.; Williams, C. M. Reassignments and Corroborations of Oxo-Bridged Natural Products Directed by OSE and DU8+ NMR Computation. *Angew. Chem.* **2019**, *58* (21), 7107–7112. https://doi.org/10.1002/ANIE.201902777.

(12) Sarotti, A. M. In Silico Reassignment of (+)-Diplopyrone by NMR Calculations: Use of a DP4/ J-DP4/DP4+/DIP Tandem to Revise Both Relative and Absolute Configuration. *J.*

Org. Chem. **2020**, *85* (17), 11566–11570. https://doi.org/10.1021/ACS.JOC.0C01563/ASSET/IMAGES/LARGE/JO0C01563_0005.JPEG.

(13) Kutateladze, A. G.; Reddy, D. S. High-Throughput in Silico Structure Validation and Revision of Halogenated Natural Products Is Enabled by Parametric Corrections to DFT-Computed 13C NMR Chemical Shifts and Spin-Spin Coupling Constants. *J. Org. Chem.* **2017**, *82* (7), 3368–3381. https://doi.org/10.1021/ACS.JOC.7B00188/ASSET/IMAGES/MEDIUM/JO-2017-00188J_0016.GIF.

(14) Kutateladze, A. G.; Mukhina, O. A. Minimalist Relativistic Force Field: Prediction of Proton-Proton Coupling Constants in 1H NMR Spectra Is Perfected with NBO Hybridization Parameters. *J. Org. Chem.* **2015**, *80* (10), 5218–5225. https://doi.org/10.1021/ACS.JOC.5B00619/SUPPL_FILE/JO5B00619_SI_001.PDF.

(15) Smith, S. G.; Goodman, J. M. Assigning Stereochemistry to Single Diastereoisomers by GIAO NMR Calculation: The DP4 Probability. *J. Am. Chem. Soc.* **2010**, *132* (37), 12946–12959. https://doi.org/10.1021/JA105035R/SUPPL_FILE/JA105035R_SI_001.PDF.

(16) Ermanis, K.; Parkes, K. E. B.; Agback, T.; Goodman, J. M. Doubling the Power of DP4 for Computational Structure Elucidation. *Org. Biomol. Chem.* **2017**, *15* (42), 8998–9007. https://doi.org/10.1039/C7OB01379E.

(17) Howarth, A.; Ermanis, K.; Goodman, J. M. DP4-AI Automated NMR Data Analysis: Straight from Spectrometer to Structure. *Chem. Sci.* **2020**, *11* (17), 4351–4359. https://doi.org/10.1039/D0SC00442A.

(18) Grimblat, N.; Zanardi, M. M.; Sarotti, A. M. Beyond DP4: An Improved Probability for the Stereochemical Assignment of Isomeric Compounds Using Quantum Chemical Calculations of NMR Shifts. *J. Org. Chem.* **2015**, *80* (24), 12526–12534. https://doi.org/10.1021/ACS.JOC.5B02396/SUPPL_FILE/JO5B02396_SI_002.XLSX.

(19) Grimblat, N.; Gavín, J. A.; Hernández Daranas, A.; Sarotti, A. M. Combining the Power of J Coupling and DP4 Analysis on Stereochemical Assignments: The J-DP4 Methods. *Org. Lett.* **2019**, *21* (11), 4003–4007. https://doi.org/10.1021/ACS.ORGLETT.9B01193/SUPPL_FILE/OL9B01193_SI_001.PDF.

(20) Howarth, A.; Goodman, J. M. The DP5 Probability, Quantification and Visualisation of Structural Uncertainty in Single Molecules. *Chem. Sci.* **2022**, *13* (12), 3507–3518. https://doi.org/10.1039/D1SC04406K.

(21) Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J. Computational Prediction of 1H and 13C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem Rev* **2012**, *112* (3), 1839–1862. https://doi.org/10.1021/CR200106V/ASSET/IMAGES/MEDIUM/CR-2011-00106V_0009.GIF.

(22) Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y.; Nannapaneni, Y.; Duggan, B. M.; Min, J.; Lin, E. C.; Gerwick, E. C.; Cottrell, G. W.; Gerwick, W. H. Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Sci. Rep. 2017 7:1* **2017**, *7* (1), 1–17. https://doi.org/10.1038/s41598-017-13923-x.

(23) Reher, R.; Kim, H. W.; Zhang, C.; Mao, H. H.; Wang, M.; Nothias, L. F.; Caraballo-Rodriguez, A. M.; Glukhov, E.; Teke, B.; Leao, T.; Alexander, K. L.; Duggan, B. M.; Van Everbroeck, E. L.; Dorrestein, P. C.; Cottrell, G. W.; Gerwick, W. H. A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products. *J. Am. Chem. Soc.* **2020**, *142* (9), 4114–4120. https://doi.org/10.1021/JACS.9B13786/SUPPL_FILE/JA9B13786_SI_002.ZIP.

(24) Kim, H. W.; Zhang, C.; Reher, R.; Wang, M.; Alexander, K. L.; Nothias, L. F.; Han, Y. K.; Shin, H.; Lee, K. Y.; Lee, K. H.; Kim, M. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. DeepSAT: Learning Molecular Structures from Nuclear Magnetic Resonance Data. *J Cheminform* **2023**, *15* (1), 1–12. https://doi.org/10.1186/S13321-023-00738-4/FIGURES/3.

(25) Zanardi, M. M.; Sarotti, A. M. GIAO C-H COSY Simulations Merged with Artificial Neural Networks Pattern Recognition Analysis. Pushing the Structural Validation a Step Forward. *J. Org. Chem.* **2015**, *80* (19), 9371–9378. https://doi.org/10.1021/ACS.JOC.5B01663/ASSET/IMAGES/JO-2015-01663B_M002.GIF.

(26) Hinneburg, A.; Bj¨, B.; Egert, B.; Porzel, A. Duplicate Detection of 2D-NMR Spectra. *J. Integr. Bioinform.* **2007**, *4* (1), 64–80. https://doi.org/10.1515/JIB-2007-53.

(27) Bodis, L.; Ross, A.; Bodis, J.; Pretsch, E. Automatic Compatibility Tests of HSQC NMR Spectra with Proposed Structures of Chemical Compounds. *Talanta* **2009**, *79* (5), 1379–1386. https://doi.org/10.1016/J.TALANTA.2009.06.017.

(28) Pierens, G. K.; Mobli, M.; Vegh, V. Effective Protocol for Database Similarity Searching of Heteronuclear Single Quantum Coherence Spectra. *Anal. Chem.* **2009**, *81* (22), 9329–9335. https://doi.org/10.1021/AC901616T/SUPPL_FILE/AC901616T_SI_001.PDF.

(29) Yang, Z.; Vegh, V.; Reutens, D. C.; Pierens, G. K. A Rapid Procedure for Spectral Similarity Matching of Heteronuclear Single Quantum Coherence Spectra. *DICTA 2011* **2011**, 302–307.

(30) Han, J.; Kang, H.; Kang, S.; Kwon, Y.; Lee, D.; Choi, Y. S. Scalable Graph Neural Network for NMR Chemical Shift Prediction. *Phys. Chem. Chem. Phys.* **2022**, *24* (43), 26870–26878. https://doi.org/10.1039/D2CP04542G.

(31) Guan, Y.; Shree Sowndarya, S. V.; Gallegos, L. C.; St. John, P. C.; Paton, R. S. Real-Time Prediction of 1H and 13C Chemical Shifts with DFT Accuracy Using a 3D Graph Neural Network. *Chem. Sci.* **2021**, *12* (36), 12012–12026. https://doi.org/10.1039/D1SC03343C.

(32) Kwon, Y.; Lee, D.; Choi, Y. S.; Kang, M.; Kang, S. Neural Message Passing for NMR Chemical Shift Prediction. *J. Chem. Inf. Model.* **2020**, *60* (4), 2024–2030.

https://doi.org/10.1021/ACS.JCIM.0C00195/ASSET/IMAGES/LARGE/CI0C00195_0003.JPEG.

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 16 [Computer Software]. Gaussian, Inc. Retrieved from https://www.gaussian.com. 2016. **2016**.

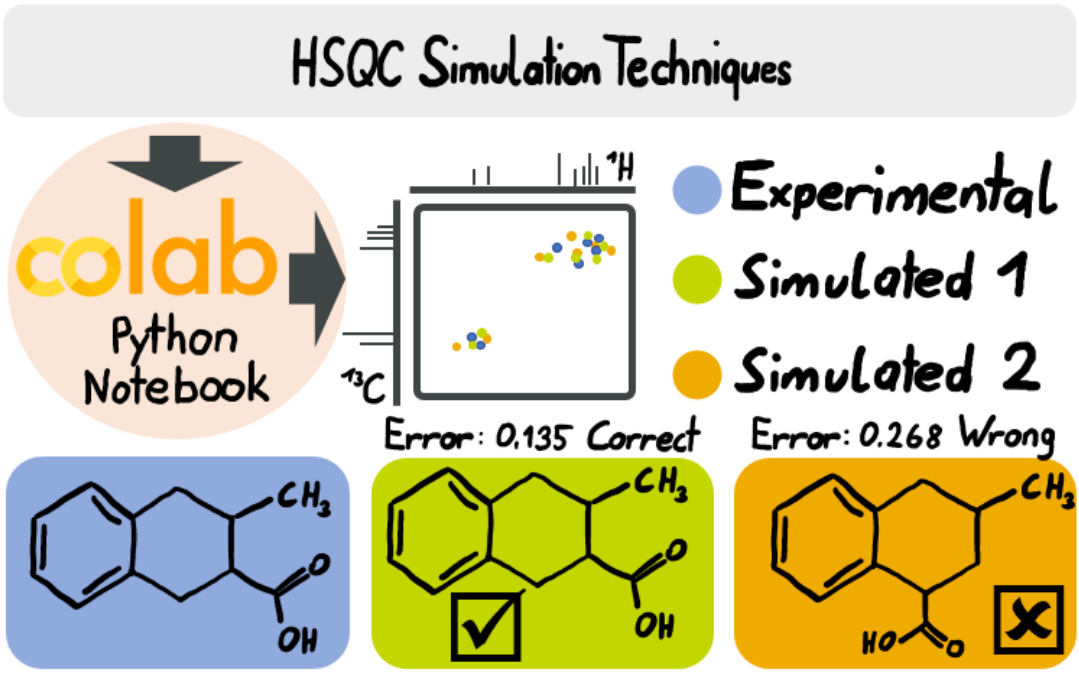
(34) Wishart, D. S.; Guo, A. C.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, H. B.; Greiner, R.; Gautam, V. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res* **2022**, *50* (D1), D622–D631. https://doi.org/10.1093/NAR/GKAB1062.

(35) Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist. Q.* **1955**, *2* (1–2), 83–97. https://doi.org/10.1002/NAV.3800020109.

(36) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (4), 747–750. https://doi.org/10.1021/CI9803381/ASSET/IMAGES/MEDIUM/CI9803381E00016.GIF.

(37) Wu, F.; Radev, D.; Li, S. Z. Molformer: Motif-Based Transformer on 3D Heterogeneous Molecular Graphs. *arXiv:2110.01191* **2021**.

Graphical Abstract:



A Google Colab tool was developed that utilizes Heteronuclear Single Quantum Coherence (HSQC) spectra simulations for compound identification and assignment by comparing two simulated spectra with experimental data, and selecting the likelier correct compound based on matching error calculation.