(Semi-) Automatic Review Process for Common Compound Characterization Data in Organic Synthesis

Yu-Chieh Huang[a], *Pierre Tremouilhac*[a], *Stefan Kuhn*[c], *Pei-chi Huang*[a], *Chia-Lin Lin*[a], *Nils Schlörer*[e], *Oskar Taubert*[d], *Markus Götz*[d,f], *Nicole Jung**[a,b], *Stefan Bräse**[a,b]

Email: yu-chieh.huang@kit.edu; stefan.kuhn@ut.ee; nicole.jung@kit.edu; pierre.tremouilhac@kit.edu; stefan.braese@kit.edu: nils.schloerer@uni-koeln.de; oskar.taubert@kit.edu; markus.goetz@kit.edu

[a]Institute of Biological and Chemical Systems (IBCS-FMS), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe Germany; [b]Institute of Organic Chemistry, Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany; [c]University of Tartu, Institute of Computer Science, Narva mnt 18, 51009 Tartu, Tartumaa, Estonia; [d]Scientific Computing Center (SCC), Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany; [e]Faculty of Chemistry and Earth Sciences, NMR platform, Humboldtstraße 10, 07743 Jena, Germany; [f]Helmholtz AI, Germany.

ABSTRACT

A method for data review in chemical sciences with a focus on data for the characterization of synthetic molecules is described. As current procedures for data curation in chemistry rely almost exclusively on manual checking or peer reviewing, a (semi-)automatic procedure for the evaluation of data assigned to molecular structures is proposed and demonstrated. The information usually required for the identification of isolated compounds is used to clarify whether the data is complete with respect to the available data types and metadata, if it is consistent with the proposed structure and if it is plausible in comparison to simulated data. Spectra prediction and automatic signal comparison are applied to NMR evaluation, mass spectrometry data are evaluated by signal extraction, and machine learning is used for IR analysis. The proposed protocol shows how an integration of different tools for data analysis can help to overcome the challenges of the currently purely manual reviewing and curation efforts for data in synthetic chemistry.

Keywords: data curation, repositories, electronic lab notebooks, chemistry data, analytics

BACKGROUND

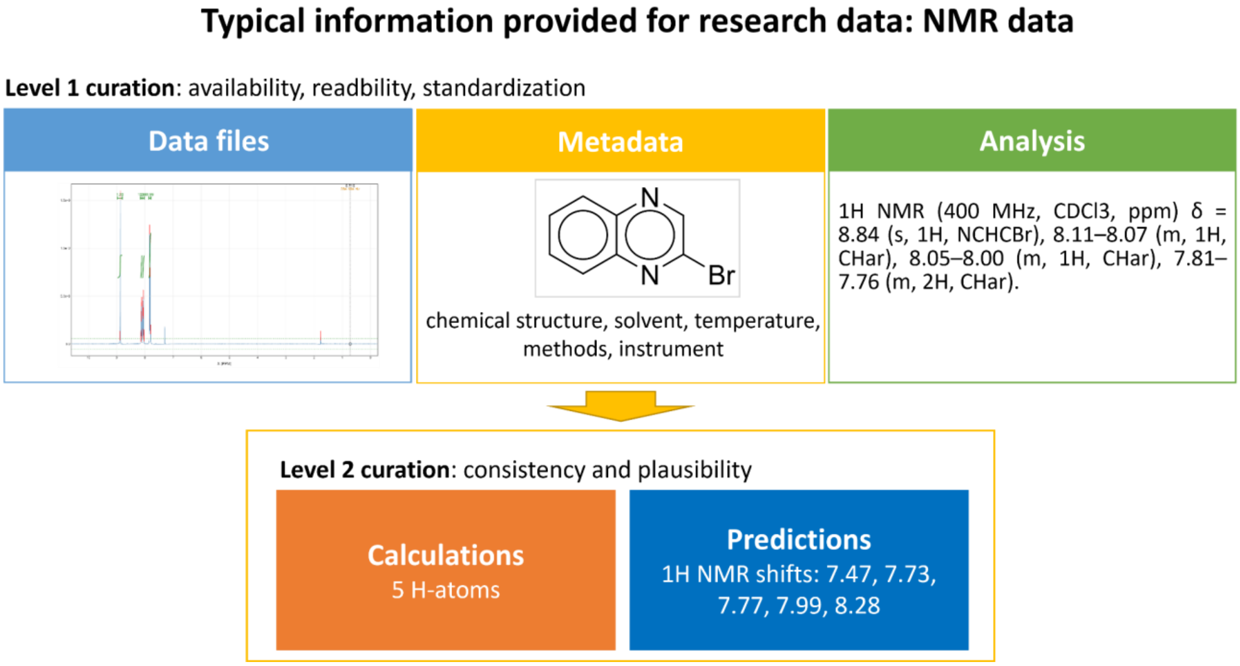
Research data play an essential role in all scientific disciplines as evidence of the research results is obtained. They also serve as an important source of information in various re-use scenarios, including the replication of results, a comparison with other datasets, or analytical purposes such as machine learning (ML). Research data can advance scientific work and their preservation can accelerate the gain of knowledge. In chemistry, as in many other experimental disciplines, measurement data are particularly important. Their storage, preservation, and reuse can take place at different stages of the research process and, depending on this, can be facilitated either by electronic laboratory journals, databases, and institutional or non-institutional repositories. In any case, when dealing with research data, the question arises as to how the data provided can be curated to ensure compliance with either formal or content-related needs. While formal requirements can usually be demanded or even enforced quite easily, checking the content of research data is a major challenge. Reasons for this are, firstly, the complexity of the research and the accompanying diversity of research data and their analyses and, secondly, the lack of standardisation and open data formats in many areas. Therefore, currently, the formal curation and content review of most of the research data – if carried out at all – is based on manual peer-reviewing. Only some established infrastructures such as the Cambridge Crystallographic Data Centre (CCDC)[1] were able to implement suitable processes for an automated evaluation of research data, which allow for the efficient and automatic review of thousands of crystal structures for the Crystal Structure Database (CSD) per year[2,3]. While the CCDC developed advanced functions for crystal structures, automated methods for data curation and review are becoming increasingly important also for other data types as well. In the future, a significant increase in the utilisation of repositories and databases can be anticipated due to an increased obligation by publishers[4,5,6,7] and funding organisations[8,9,10,11,12] to preserve and provide access to research data. Review and curation mechanisms could also provide valuable assistance for other services than repositories. For example, the software solutions used by researchers for documentation, above all electronic laboratory journals, could benefit from support through automatic review mechanisms.

PREVIOUS WORK

In the past, various algorithms, models, software tools, and web-services have been described that can be used for efficient, (partially) automated processing and curation of data in the discipline of synthetic chemistry. The necessary input for those systems is generally the chemical structure in combination with either the data file or the textual description (hereinafter referred to as named “analysis”) of the measurement. For NMR measurements of organic molecules, there are several established rule-based systems available. nmrshiftdb2 offers a web service that can be used to compare NMR shifts of experimentally gained ^{1}H NMR and ^{13}C NMR data with chemical shifts that are simulated based on the chemical structure of the expected molecule.[13] nmrshiftdb2 facilitates quality control for spectra assignments of small organic molecules and supports scientists to approve their structures.[14] Another web service, CSEARCH[15], offers functions to check proposed chemical structures by comparison of the experimentally found ^{13}C NMR data with the respective calculated data. Important tools that can aid curation efforts are also provided by so-called CASE software (computer-aided structure elucidation)[16] such as seneca[17] or LSD[18]. The web-based software CASPER[19,20] (computer-assisted spectrum evaluation of regular polysaccharides) uses NMR measurements to elucidate carbohydrates, including new oligosaccharides and glycoconjugates, based on the agreement with predicted ^{1}H NMR and ^{13}C NMR chemical shifts. Other models such as "*Ask Ernö*"[21] work independently of selected compound classes, offering self-learning systems for the automatic analysis of ^{1}H and ^{13}C NMR spectra based on experimental data. ML techniques have been applied to NMR signal processing, prediction, and structure verification.[22,23] Several approaches use DFT-computed data as input for the training of (deep) neural networks gaining suitable models to predict ^{13}C NMR shifts to e.g. identify, structural mis-assignments of organic compounds.[24,25] Other work with neural networks has accelerated the development of suitable simulation methods for ^{1}H NMR and ^{13}C NMR-based shifts extracted from experimental data in the last years.[26, 27, 28, 29, 30, 31, 32, 33, 34,35] For the simulation of IR spectroscopic data, DFT calculations can be used for specific scientific challenges and general approaches were described in previous work.[36] However, several ML approaches for the simulation of IR-vibrational data were also reported.[37,38, 39, 40, 41] Also web services were built to facilitate the comparison of experimental IR data with previously reported data[42] and the analysis of given spectra[43] and the prediction of spectra from a given structure.[, 44, 45, 46]

IMPLEMENTATION

For a review of research data, various types of information can be taken into account. Measurement data consist primarily of (a) data files such as spectroscopic data from devices including data and metadata, (b) additional associated metadata and descriptions such as the structure of the measured substance or the measurement parameters, and (c) the interpretation of the data of the measured samples in textual form (analysis)[47]. Special options result from the chemical structure, which enable the calculation or prediction/simulation of characteristic parameters for a measured substance (Figure 1). The systematic description of the types of information (a) - (c) shown in Figure 1 can be transferred to a wide variety of measurement data. For the characterisation of substances in synthetic organic chemistry, these are NMR spectroscopy, mass spectrometry, IR or UV-Vis spectroscopy data, elemental analysis, and several others.


Typical information provided for research data: NMR data
Level 1 curation: availability, readbility, standardization
Data files
Metadata
N
N
Br
chemical structure, solvent, temperature, methods, instrument
Analysis
Level 2 curation: consistency and plausibility
Calculations
5 H-atoms
Predictions
1H NMR shifts: 7.47, 7.73, 7.77, 7.99, 8.28

1H NMR (400 MHz, CDCl3, ppm) δ = 8.84 (s, 1H, NCHCBr), 8.11–8.07 (m, 1H, CHar), 8.05–8.00 (m, 1H, CHar), 7.81–7.76 (m, 2H, CHar).

Figure 1. Typical NMR information for the characterization of molecules. The characterisation is gained by several measurements each of which consists of a data file, metadata, and the analysis. Based on the information in the metadata, additional information can be gained by calculations and predictions.

The process described here uses the most frequently described data types belonging to the standard techniques for the characterisation of organic compounds - which are ^{1}H NMR, ^{13}C NMR, IR spectroscopy, and mass spectrometry. It checks those data types for the three different aspects of readability, consistency, and plausibility (Figure 2) and combines the results in a numerical value to distinguish data of high quality from data with lower quality. While we use different processes for each of the data types, there are some general principles that can be followed: Data readability and other indicators directly accessible from data files can be evaluated by checking the data for the presence of open data file formats and their validation if standardised data exists. For automatic checking of data consistency and plausibility, readable or better machine readable, standardised data are a prerequisite. Information on data consistency can be gained e.g. from the comparison of the calculated key indicators of a molecule –such as the number of atoms– with the indicators given in the analysis of data. Data plausibility is obtained from a comparison of the contents of analytical data files with the predicted data that can be obtained from the processing and simulation of molecular structures.

The implementation of the proposed process was carried out in the Chemotion ELN (ELN = Electronic Lab Notebook)[48] as well as the Chemotion repository[49,50]. The systems were chosen to evaluate different use cases for semi-automated data curation. The implementation of the process within the Chemotion ELN offers data curation features in the form of recommendations for scientists which may prevent errors. The implementation in the Chemotion repository can serve as a recommender for data providers and as a curation tool for reviewers.

Depending on how detailed research data are to be curated, our processes define three levels: Level 1 summarises the methods that perform a check of data files, metadata, and analyses without further information („one-dimensional evaluation“), level 2 consists of comparisons of data with calculated or simulated data („two-dimensional evaluation“), and level 3 combines the results of level 1 and level 2.

Data Availability checks within Level 1

The procedure for checking data within level 1 consists of the check of data for their availability, meaning the existence of a data file, its readability/processability and standardisation (Figure 2). Further, the availability of a machine-readable chemical structure is checked. The machine-readable chemical structure can be considered as additional metadata. The availability of analyses is approved by the presence of textual descriptions for the different measurement types. For the mandatory additional metadata as well as the analyses, no further checks of readability and standardisation are necessary as the input options of the Chemotion systems enforce the support of the desired standards.

Data consistency and plausibility: Level 2 evaluation

Level 2 („two-dimensional evaluation“) requires additional information from calculated or simulated data to be used for comparisons. Typical examples are consistency checks with calculated data or plausibility checks using simulated data. In both cases, the information obtained from data files and analyses is compared to theoretical information gained from the chemical structure (Figure 2). To test the plausibility, the information obtained from the data files in the form of data points or signals is compared with the simulated data points or signals and evaluated while considering the tolerances specific to the data type. In this work, NMR and IR data are utilised for a plausibility analysis (see descriptions in the section plausibility). The consistency of the data is reviewed by comparing the data files or the data analyses with calculations that are available based on the chemical structure. Methods for NMR and mass data can be proposed to confirm the consistency of the data (see descriptions in the section consistency).

Data Consistency as part of Level 2 evaluation

In our approach, data consistency is checked for the analytical data from ^{1}H NMR, ^{13}C NMR spectroscopy and mass spectrometry completely automatically. The analytical interpretation (analysis) given by the data creator is compared with the information from automatically processed molecular structures by cheminformatics tools (cheminformatics toolkits used in this work are described in the SI). For ^{1}H NMR and ^{13}C NMR data, the signal interpretation (analysis), either extracted from data files or added manually by the scientists, is parsed to determine the total number of atoms that are referenced and the number is then compared with the molecular formula of the proposed structure (option 2, Figure 2). Regarding mass spectrometry data, the data file of the mass spectrum is checked for consistency with the expected values calculated for a specific target structure. Those procedures based on data file approval require some effort as the data may consist of diverse scans belonging to the same measurement. In the process proposed here, we established a protocol to search in all scans of one measurement for the exact mass-to-charge ratio and alternatively for the exact mass +1, +23, and +39, due to proton, sodium, and potassium adduct formation. As an additional method of ensuring consistency, the list of identified mass peaks given in the analytical information is checked for values that correspond to the exact mass of the target molecule or possible combinations thereof (option 3, Figure 2). Here again, the verification process also includes the comparison with values that correspond to the exact mass of the target molecule or combinations with the most common species derived from the molecules' exact mass (see supporting information for details) (option 4, Figure 2).

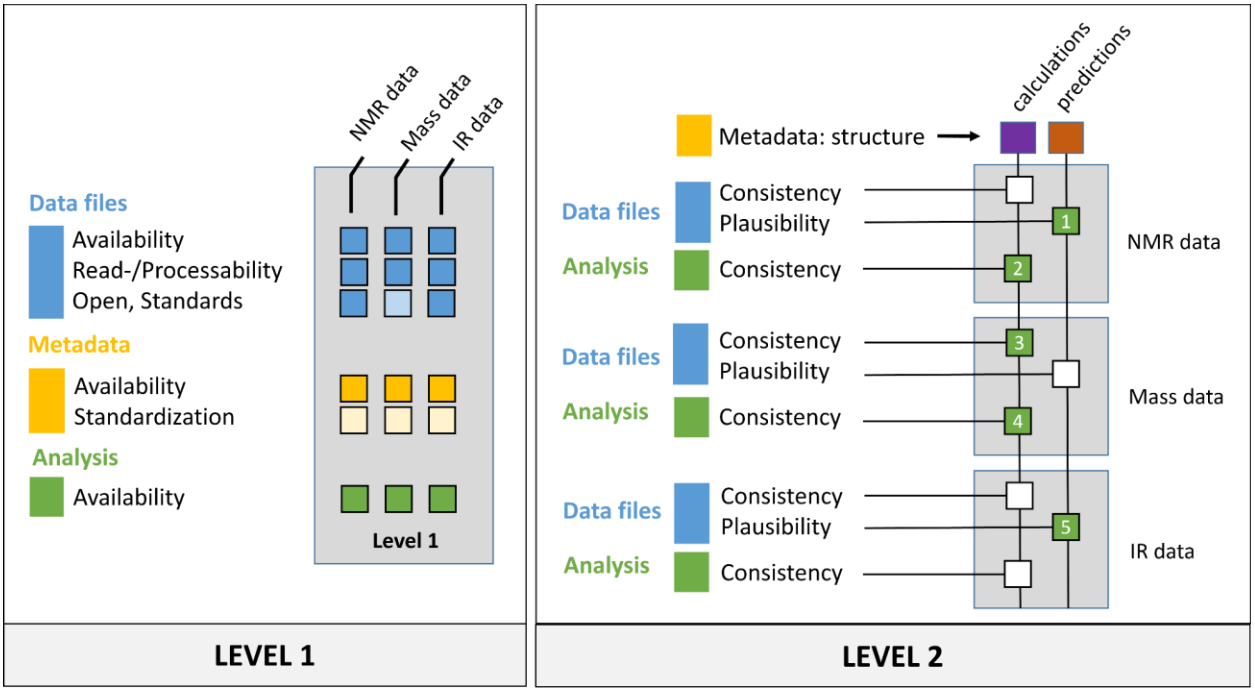
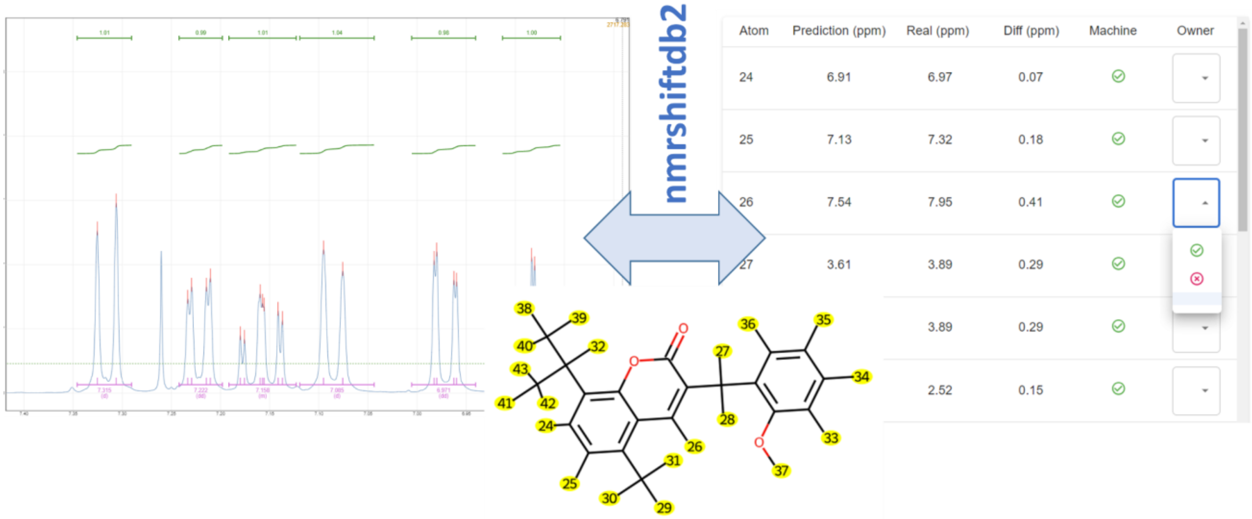


Figure 2. Aspects covered by the "one-dimensional" (left) two-dimensional (right) evaluation of data files. Level 1 includes basic checks for data, metadata and analysis for each type of measurement. Dark blue/yellow/green = good coverage, light blue/yellow/green = only partly covered due to missing standardisation or not full coverage of all information by distinct standards. Level 2 includes the evaluation of data files and analyses for consistency and plausibility with calculations and simulations resulting from the chemical structure. Green squares = evaluation is part of this work, white square = evaluation is not included. Numbers 1-5 refer to the options described in the main text.

Data Plausibility as part of Level 2 evaluation

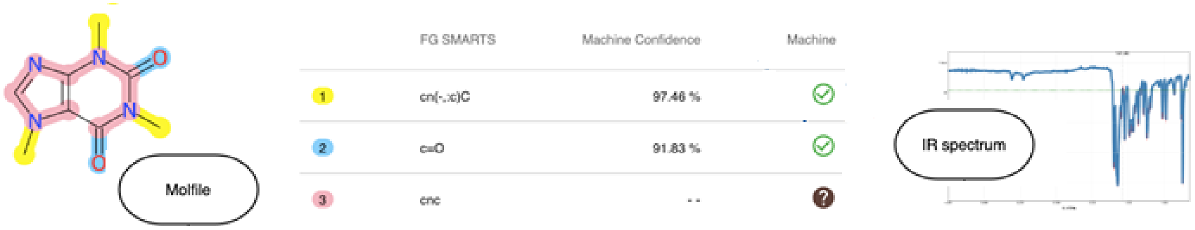
While the theoretical values necessary for the consistency checks can be gained by different cheminformatics toolkits without large differences in the outcome, the choice of simulation tool for plausibility checks is crucial for the results. In our work, testing plausibility relies on different, analytical method-specific models to predict the properties of molecules and their spectroscopic characteristics for NMR and IR spectroscopy. To check the plausibility of NMR spectroscopic data, the data files of ^{1}H NMR and ^{13}C NMR measurements are processed via the software ChemSpectra[51] and analysed using the QuickCheck service from nmrshiftdb2 (option 1, Figure 2)[14]. The process differs for ^{1}H NMR data and ^{13}C NMR data as preparation of ^{1}H NMR data as well as its analysis is more complex than the preparation and analysis of ^{13}C NMR data. In our model, ^{1}H NMR spectra to be curated must be manually annotated with regards to multiplicity assignment, and signals not belonging to the expected molecule (such as solvents and impurities) have to be removed. For ^{13}C NMR data, a process has been implemented, which allows for the automatic selection of signals from a given NMR data file (for a detailed description of the process and its limitations, see supporting information Chapter 2). The shifts of all manually and automatically selected signals from ^{1}H NMR and ^{13}C NMR data are then compared to the shifts that can be predicted for the expected molecule by the service from nmrshiftdb2. Depending on the difference between the experimentally found and the simulated shift, a status of "accept", "warning" or "reject" is assigned to each shift (Figure 3). The evaluation routine enables the manual correction of those results as the simulated data may contain errors or might not be as precise as necessary for the included tolerances. The correction necessitates manual alteration of the review outcome via confirmation of individual signals and their matching and is then included in the outcome of the evaluation.



Atom	Prediction (ppm)	Real (ppm)	Diff (ppm)	Machine	Owner
24	6.91	6.97	0.07	✓	▾
25	7.13	7.32	0.18	✓	▾
26	7.54	7.95	0.41	✓	▴ ✓ ⊗
27	3.61	3.89	0.29	✓	
		3.89	0.29	✓	▾
		2.52	0.15	✓	▾

Figure 3. Description of ^{1}H NMR plausibility checks for level 2 with the use of ^{1}H NMR data files in combination with the chemical structure of a chemical compound. After multiplicity detection for the relevant signals, the web services of nmrshiftdb2 are used to simulate the expected chemical shifts for a selected chemical structure and to add a quick interpretation of the fitness of the experimentally found (real) to the simulated shifts (predicted).

Plausibility checks for infrared spectroscopy are built in such a way as to check if those functional groups, which are part of the chemical target structure, can be identified in a provided IR spectrum (option 5, Figure 2). Our evaluation process utilises available cheminformatics toolkits such as nmrglue[52], rdkit[53] and functional group finder (IFG)[54] in combination with an ML model. The implemented model adopts convolutional neural networks to recognise the presence of functional groups from the full spectrum profile, not discrete peaks. This method is data-driven without pre-encoded rules (see also results section). The trained model is used to estimate whether IR data files contain the signals expected due to the functional groups of the corresponding chemical structure. A detailed description of the methods and results for the implementation of the model is given in the Supporting Information (starting with Chapter 7). Based on this model, the functional groups of the molecules to be curated are extracted from the given Molfile, and spectroscopic data points are extracted and preprocessed using the IR JCAMP-DX file. For each functional group that is expected, the outcome of the ML model in the form of the probability (does the spectrum reveal the presence of a functional group?) as well as the typical confidence of the model with respect to the functional group (referring to the test data) is given. Based on a combination of these two indicators, an assessment of the matching of experimental and predicted data as an indicator for the plausibility of the data is made (Figure 4).



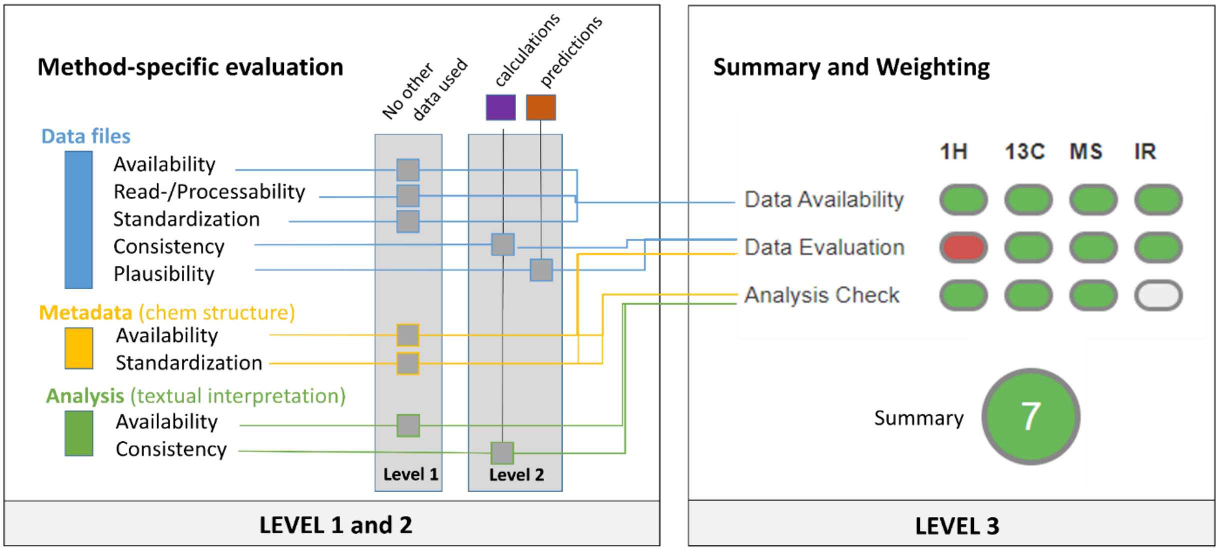
	FG SMARTS	Machine Confidence	Machine
1	cn(-,:c)C	97.46 %	✓
2	c=O	91.83 %	✓
3	cnc	- -	?

Figure 4. Schematic description of the evaluation process of IR data to determine the plausibility of the data. The chemical structure of a compound is used to extract the included functional groups. The implemented ML model gives an indication of whether the IR spectrum contains the predicted signals for the extracted functional groups. Depending on the confidence of the model for a certain functional group and the machine's result, the curator can follow the evaluation or adapt the outcome manually.

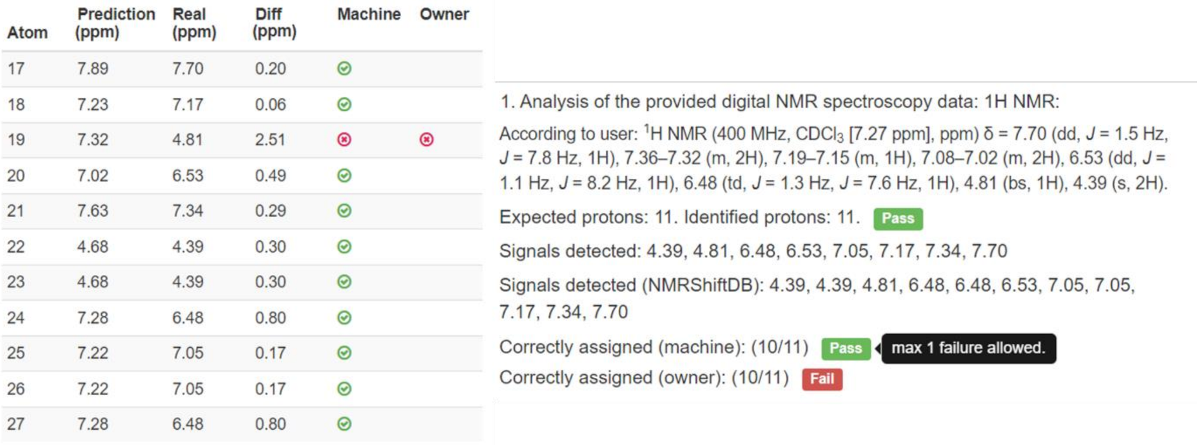
Final evaluation in Level 3

After passing level 1 and level 2 evaluations, our process combines the results from the various techniques of both levels in a level 3. For a level 3 evaluation, the individual results from the analytical techniques are weighted according to their significance (for the characterisation of a compound) and prediction accuracy (according to known strengths and weaknesses of a specific evaluation method). Level 3 gives the result of the evaluation as a combination of readability, consistency, and plausibility and is used for an estimation if the provided data are trustworthy and coherent, aiming in particular to evaluate if the proposed chemical structures match the recorded experimental data. A model for such a combined evaluation is described as level 3 of our data curation process. An example of a level 3 evaluation was obtained in the Chemotion systems ELN and repository and is depicted in Figure 5. A first overview of all results is gained via the summary of all 1st and 2nd level results. The summary is presented in three main categories referred to as "data availability", "data evaluation" and "analysis check". Each category includes various aspects of the evaluation processes of level 1 and level 2, facilitating the review of the provided data with a different focus. The availability of data files, metadata and analysis is a prerequisite for the overall process and is implemented as mandatory information to be given during the provision of the data. Therefore, the availability of data files, metadata and analysis is implicitly included in all review categories and not listed separately. The category "data availability" summarises the availability, readability/processability as well as openness/standardisation of the provided data files. The category "data evaluation" covers the results of the consistency and plausibility checks that were done based on the data files that were provided. As these results were gained based on the chemical structure and other metadata as necessary information, this category also includes the indirect check of the most important metadata availability and its standardisation. The category "analysis check" summarises the results that are obtained from comparing the provided data with calculated data. The results of the review are given as a colour code with green = passed, red = not passed, black = not available/processable and grey = not reviewed (Figure 5). For the categories "data availability" and "analysis check" the colour code is a direct result of the evaluation results, as the outcome of the evaluation is a clear true or false indication. For the category "data evaluation", a small tolerance is included for simulation-based plausibility evaluation. This should reflect that simulations may contain several weaknesses (e.g. imprecise prediction of chemical shifts in NMR spectroscopy) and do not cover all aspects of necessary information (missing confidence for functional group prediction in IR simulation). The contributing aspects for a decision on the colour-coded review are available via a summary of the most important facts on which the result is based. The details also include information on the differences between expected and obtained results and, for transparency reasons, the corrections that were made by humans to revise the weaknesses of the simulation models if there are any.

A.



B.



Atom	Prediction (ppm)	Real (ppm)	Diff (ppm)	Machine	Owner
17	7.89	7.70	0.20	✓	
18	7.23	7.17	0.06	✓	
19	7.32	4.81	2.51	⊗	⊗
20	7.02	6.53	0.49	✓	
21	7.63	7.34	0.29	✓	
22	4.68	4.39	0.30	✓	
23	4.68	4.39	0.30	✓	
24	7.28	6.48	0.80	✓	
25	7.22	7.05	0.17	✓	
26	7.22	7.05	0.17	✓	
27	7.28	6.48	0.80	✓	

1. Analysis of the provided digital NMR spectroscopy data: 1H NMR:

According to user: ^{1}H NMR (400 MHz, $CDCl_3$ [7.27 ppm], ppm) δ = 7.70 (dd, J = 1.5 Hz, J = 7.8 Hz, 1H), 7.36–7.32 (m, 2H), 7.19–7.15 (m, 1H), 7.08–7.02 (m, 2H), 6.53 (dd, J = 1.1 Hz, J = 8.2 Hz, 1H), 6.48 (td, J = 1.3 Hz, J = 7.6 Hz, 1H), 4.81 (bs, 1H), 4.39 (s, 2H).

Expected protons: 11. Identified protons: 11. Pass

Signals detected: 4.39, 4.81, 6.48, 6.53, 7.05, 7.17, 7.34, 7.70

Signals detected (NMRShiftDB): 4.39, 4.39, 4.81, 6.48, 6.48, 6.53, 7.05, 7.05, 7.17, 7.34, 7.70

Correctly assigned (machine): (10/11) Pass max 1 failure allowed.

Correctly assigned (owner): (10/11) Fail

Figure 5. A. Schematic descriptions of the dependencies of the level 3 review on the outcome of level 1 and level 2 evaluations. Arrows in different colours indicate the influence of distinct level 1 and 2 evaluations on different review categories in level 3. Summary and weighting are described as a one number indication, considering the results from all types of measurements (NMR, Mass, and IR, for details of the process, see supplemental information). **B**. Example for a typical level 2 evaluation of 1H NMR data in detail. The consistency check refers to the comparison of counted protons in the textual interpretation (analysis) with the calculated number of protons according to the chemical structure. The plausibility check includes the prediction of shifts for the chemical structure (with the service of nmrshiftdb2) and the comparison to the real values as extracted from the NMR data files. The example results in 10/11 correctly assigned shifts, one shift needs manual review by the curator.

For a review of data at a glance, the evaluation results – including all aspects of each analysis method – are reflected in one number ranging from -2 points (missing and probably false or contradictory information) to 10 points (data files, metadata, and analysis in full accordance with the expected outcome). This number should give a clear indication of the coherence of a dataset to be used for reviewing purposes but should also serve as a general indicator for comparing the quality of different datasets. To obtain a meaningful indicator, the results from the data tests are weighted, i.e. depending on the significance, informative value of the individual measurement method, and the accuracy of fit or susceptibility to the error of the respective test method, the individual evaluations from levels 1 and 2 are included to a different extent in the final evaluation. The rules that define the evaluation of the individual aspects and their weighting are proposed in this work as a possible catalogue that enables the evaluation of data for substance characterisation in organic chemistry. The catalogue of rules used to evaluate data according to the work discussed here is described in more detail in the Supporting Information. It includes similar principles that are also important in the evaluation of data by a peer review. Thus, a goodness of fit of the data from ^{1}H NMR and ^{13}C NMR testing is assumed to be necessary. A hard criterion here is, for example, the consistent comparison of the signals found in the NMR evaluation with the number of signals to be expected. An equally important aspect is finding the correct molecular mass (or an acceptable adduct/fragment) in both the evaluation of the data and the identification in the data file provided. A lower weighting is assigned to the accuracy of fit of simulated data and experimentally found data in the data file. A distinction is also made for the available types of analysis: ^{13}C NMR experiments can usually be simulated well and also in many cases the data files can be used for comparisons through simple, automated processing. ^{1}H NMR data are somewhat more difficult to incorporate into an automated workflow and the simulations are usually more complex, so ^{1}H NMR simulations are weighted less highly. The accuracy of the fit of the IR simulations to the experimental data is also given little weight in the evaluation.

RESULTS

The usefulness of the described system depends on a meaningful outcome but also on the effort that has to be invested to achieve the intended outcome. A low reviewing effort compared to a traditional review of data is a key requirement to partly replace the current processes in research data repositories or other management systems for research data. Consequently, concepts to automate the described curation processes were integrated for all suitable types of measurements and data. According to our findings, all evaluated data types are generally suitable for automated curation with respect to data availability, readability, standardisation, consistency, and plausibility with limitations particularly for the curation with respect to ^{1}H NMR plausibility. For the latter, the system requires the support of the reviewer in those cases where multiplicity selection and integration are missing (Figure 6). This limitation referring to ^{1}H NMR spectra is due to the complexity of including overlapping signals which make a standardised and automated analysis hard to achieve.

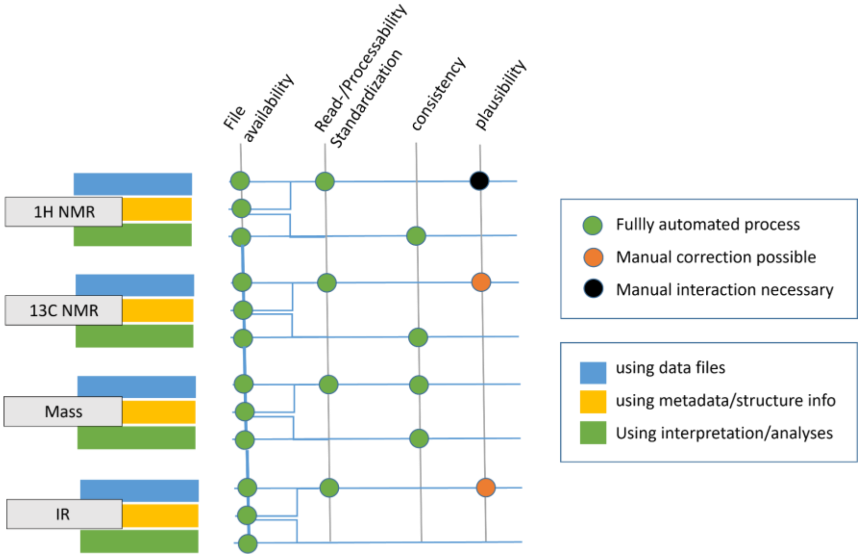

File availability
Read-/Processability Standardization
consistency
plausibility
1H NMR
13C NMR
Mass
IR
Fullly automated process
Manual correction possible
Manual interaction necessary
using data files
using metadata/structure info
Using interpretation/analyses

Figure 6. Evaluation processes and their degree of recommended automation for the different measurements included in the review process.

Evaluation by testing and approval of the process in the Chemotion repository

For an evaluation of the implemented processes, 110 exemplarily taken datasets published in the Chemotion repository by the scientists who synthesised the compounds were curated according to the processes described in the chapter implementation. The outcome of this half-automated curation was then evaluated by human curators. All exemplarily used datasets consist of spectroscopic data of organic molecules, for which ^{1}H NMR, ^{13}C NMR, mass and IR data are available in a machine_readable format. Datasets with obvious errors such as missing signals that are already indicated by the data provider or datasets belonging to structures that cannot be evaluated due to the absence of prediction tools, such as organometallic compounds, were excluded from the chosen exemplarily dataset (see SI Chapter 7 (3) for further details). For all data, the curator needed to (1) open the dataset for the sample, (2) initiate the automatic simulation of ^{1}H NMR and ^{13}C NMR, (3) obtain a review summary upon clicking. In some cases, where the submitter of the data did not add any multiplicity information in the ^{1}H NMR spectra, this was done by the curator. Additionally, ^{13}C NMR data that needed correction due to a wrongly assigned threshold or missing multiplicity information, was adjusted. After these steps, a first scoring can be obtained, giving information on the suitability of the model without further evaluation by the curator. Scores 10 and 9 can be reached, if all data are consistent and plausible (within the accepted tolerance range, e.g. for NMR spectroscopy shifts). Score 10 can be gained if there is no difference in the data with respect to the plausibility and consistency checks, score 9 can be reached for the same with the exception of IR data mismatch. In every case where the data score did not reach 9 or 10 immediately, the curator examined the ^{1}H NMR and ^{13}C NMR data in detail to determine if the data's variations from the simulation were acceptable or if there were other justifications for the correction of the evaluation result by the human curator. From the 110 datasets (further referred to as full dataset, Figure 8, a), we identified 18 examples in the dataset for which ^{1}H and/or ^{13}C predictions were not possible because the detected amount of signals did not correspond to the number of simulated ones, therefore, the NMR plausibility check was not possible. The reasons for this incompatible match are manifold. Very often, different routines of the Chemotion repository and the used service from nmrshiftdb2 are the reason for this (see information in SI, Chapter 6.2). Removing these data, for which a simulation was not possible, and therefore an automatic comparison not available, from the evaluation routine gave a dataset of 92 examples (Figure 8, b). We found that 27 examples that correspond to 29% (reduced dataset) of the data were evaluated as fully consistent without human curation (assuming that mismatch of IR simulation data is tolerated, Figure 8, c). For a further 18 datasets, only minor differences (one signal in ^{1}H NMR or ^{13}C NMR outside the tolerance, no difference in IR and mass) were found

(Figure 8, e). The experience of the curation reveals that minor differences of the real data to the simulated data, in particular ^{1}H NMR and ^{13}C NMR data can be tolerated and, therefore, an automated process assigning high scores also to data with minor differences can still be suitable to automatically differentiate consistent data from data that needs to be checked again for further clarification. For those examples, a future version of the routine could assign a score of 9 or 10 automatically (which would increase the suitability results according to an automated process to 49% (45 datasets for the reduced dataset, Figure 8, g). Another 47 datasets (Figure 8, f) were found to have more than one mismatch in the NMR evaluation which could be clarified by examining the distinct NMR shifts in 40 of the given cases (Figure 8, f). The check of those examples with more than one non-fitting shift in the NMR data comparison can be considered a quick check. After the quick check, 85 datasets (92%) were scored with 9 or 10 (Figure 8, j). Nine remaining submissions did not fit well to the expected outcome (less than score 10 or 9 after the review of the 47 examples) and the differences of calculation/prediction to the obtained data was too large to be tolerated with a quick check of the data (Figure 8, i). The considered data was considered to be potentially wrong and required further investigation. Out of these, 7 compounds were found to be correct - they were false negative results of the curation process (Figure 8, k) and two out of nine datasets with a score from 0-7 were found to have incorrect structures assigned to the dataset. These results show that the model used is also capable of identifying errors in the assignment of data to structures. A detailed look at these data reveals how powerful the described process is: the data obtained was consistent with respect to the identified mass, therefore, mismatching datasets are considered to be most likely isomers of the target compounds which were assigned wrongly to the gained data. In the two mentioned cases, the results from the curation routine were compared with the results of possible isomers and the best-fitting result was taken as the most likely assignment of the data to the correct isomer. In both cases the results were compared with the literature and showed that the most likely assignment according to the curation process was also the commonly reported structure in the literature (see SI Chapter 6.4 for further information).

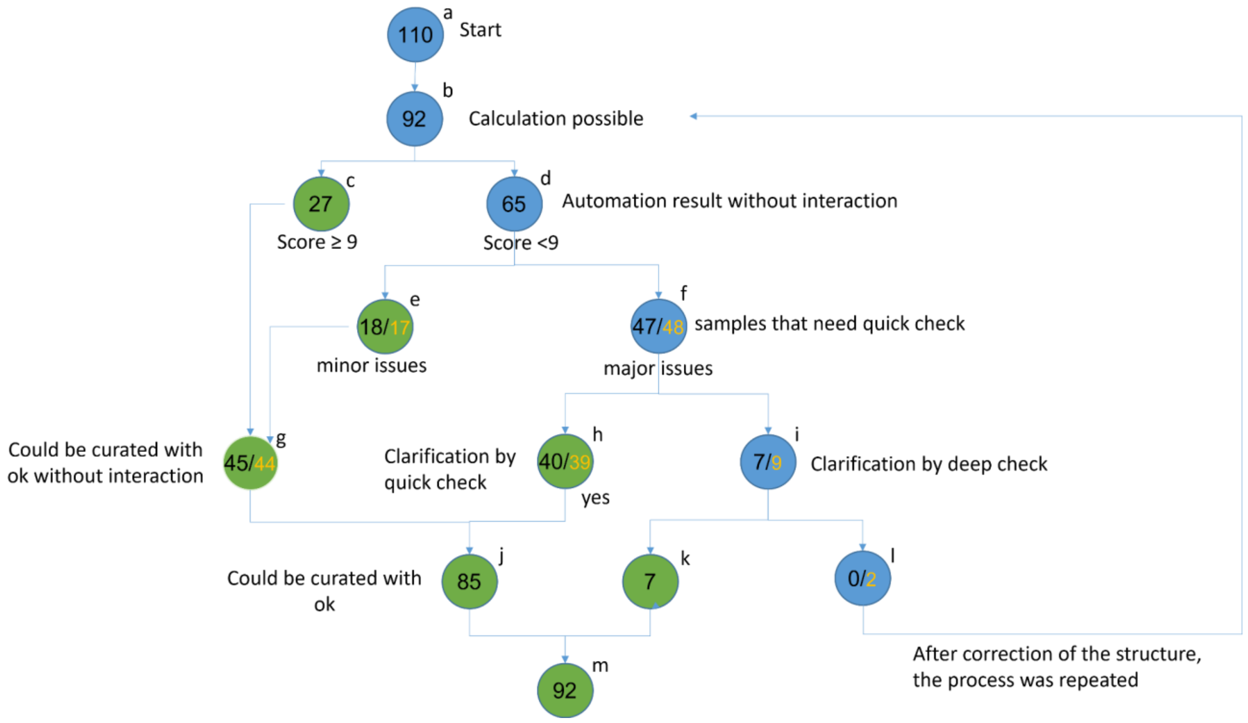

a
110
Start
b
92
Calculation possible
c
27
Score ≥ 9
d
65
Automation result without interaction
Score <9
e
18/17
minor issues
f
47/48
samples that need quick check
major issues
Could be curated with ok without interaction
g
45/44
Clarification by quick check
h
40/39
yes
i
7/9
Clarification by deep check
Could be curated with ok
j
85
k
7
l
0/2
After correction of the structure, the process was repeated
m
92

Figure 8. Outcome of the investigation of 110 datasets submitted to the Chemotion repository of which 92 could be used for half-automated curation. The effort that needed to be invested to curate the datasets was recorded for different levels of time investment. Green = number of datasets that passed the process. Yellow numbers (9 and 2 in i and l) = evaluation result before repetition of the evaluation with the corrected structures. Detailed information on the numbers is available in SI Part 2.

CONCLUSION

The presented concept shows a powerful method to support either scientists in their daily work or reviewers' work on the curation of data in the field of organic chemistry. The curation process uses a combination of different tools to check the coherence of experimental data with simulated and calculated data gained from different cheminformatics tools. It includes the analyses of ^{1}H NMR, ^{13}C NMR spectroscopy, mass spectrometry and IR spectroscopy data and a combination of the evaluation results by weighing their importance. The developed curation process was implemented in the Chemotion repository to investigate the potential benefit for data reviewers. We evaluated datasets of 110 chemical compounds provided by scientists who produced the data for publication purposes. The performance of the curation routine was described for the curator's work in general and further options to partially automate the curation routine by non-manual review processes. The implementation had a direct benefit for the reviewers as the automated checking routines and the comparison of data with simulated data could be used directly for evaluation without the need to use external software or tools. Without the additional curation of the reviewers, 49% of the data could be evaluated. With little support from the curators, 92% of the data passed the evaluation, indicating a match between the data and the proposed structures. For those examples, the curation process was improved and accelerated compared to pure manual curation. The curation process improved and accelerated the review for well-fitting data and enabled quick identification of datasets requiring a detailed review due to mismatching results. The model was shown to be suitable to identify datasets with errors, which were demonstrated by two examples where an unusual tautomer was assigned to the data. This result indicates that the tolerances of the curation routine are highly suitable to verify consistent data and detect possible errors. While the described processes facilitate the work of a data curator already in the current version, forthcoming improvements of simulation tools, particularly for NMR and IR data, could facilitate the process to a point where most of the data can be curated automatically.

ABBREVIATIONS

ELN, Electronic Laboratory Notebook; NMR, nuclear magnetic resonance; IR, infra red; DB, database; UI, User Interface; ML, Machine Learning.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The implementation of the work described above for levels 1-3 was achieved in the Chemotion ELN[55] and the Chemotion repository[56]. Both systems provide the necessary structure of elements as data input. The deep learning method for IR prediction was developed as an independent project and is available on github[57] and references on Zenodo.[58]

The SI contains a summary of all results of the curation process as a screenshot of the obtained scoring (SI Part 1) and a summary of the results as a spreadsheet (SI Part 2). Data that were used for the evaluation of the herein described model is freely available via the Chemotion repository. The DOIs to access the data are available in the supplemental Information (SI Part 2). The curation results can be directly accessed via these links.

Competing interests

The authors declare no competing interests.

Funding

This project has been funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, BR1750/34-1).

Authors' contributions

YCH designed, developed, and implemented the main architecture of the curation tool and developed the IR prediction ML model. PT, PCH, and CLL supported the embedding of the required features into the systems Chemotion ELN and Chemotion repository and adapted/maintained the work since the early developments. SK and NS supported the integration of the nmrshiftdb2 service and provided necessary adaptations. OT and MG supported the development of the ML model to predict IR data. NJ and SB contributed to the conceptual work of this project and contributed by writing the manuscript. All authors edited the manuscript.

Acknowledgements

This work was supported by the Helmholtz research field Information. We are very thankful for the support of KNMFi for hosting Chemotion at KIT and the members of the Stefan Bräse group who contributed with manifold suggestions to an ongoing improvement of the ELN. This work has benefited from the National Research Data Infrastructure for Chemistry (NFDI4Chem, project number: 441958208). We are thankful to John Jolliffe (JGU Mainz) for critical review of the manuscript. We thank Simone Gräßle, Nicolai Wippert, Laura Holzhauer, Helena Simek, Jerome Klein, Victor Larignon, Christoph Schissler, Simon Osswald, Zhen Zhang, Jasmin Busch, Florian

Mohr, Sylvia Vanderheiden, Christoph Zippel, Mareen Stahlberger, Nicolai Rosenbaum (all Karlsruhe Institute of Technology, KIT), and Fabian Fink (RWTH Aachen) for providing data to the Chemotion repository which was used for the assessment of the herein described evaluation tools.

REFERENCES

1. Groom, C. R., Allen, F. H. & Henderson, S. The Cambridge Structural Database (CSD). in *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* (Elsevier, 2019). doi:10.1016/b978-0-12-409547-2.02529-4.
2. Bruno, I. J., Shields, G. P. & Taylor, R. Deducing chemical structure from crystallographically determined atomic coordinates. *Acta Crystallogr. B* **67**, 333–349 (2011).
3. Groom, C. R. & Allen, F. H. The Cambridge Structural Database in retrospect and prospect. *Angew. Chem. Int. Ed Engl.* **53**, 662–671 (2014).
4. Hunter, A. M., Carreira, E. M. & Miller, S. J. Encouraging Submission of FAIR Data at The Journal of Organic Chemistry and Organic Letters. *Org. Lett.* **22**, 1231–1232 (2020).
5. Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. & Goudie, S. Developing a research data policy framework for all journals and publishers. *Data Sci. J.* **19**, 5 (2020).
6. *Digital Libraries: Supporting Open Science: 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 – February 1, 2019, Proceedings*. (Springer, Cham, 2019). doi:10.1007/978-3-030-11226-4.
7. Jones, L., Grant, R. & Hrynaszkiewicz, I. Implementing publisher policies that inform, support and encourage authors to share data: two case studies. *Insights Imaging* **32**, (2019).

8. *Guidelines for Safeguarding Good Research Practice: Code of Conduct*. (Deutsche Forschungsgemeinschaft, 2019).

9. EUROPEAN COMMISSION & Directorate-General for Research & Innovation. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (2017).

10. Böker, E. funder-guidelines. *Forschungsdaten.info* https://www.forschungsdaten.info/praxis-kompakt/english-pages/funder-guidelines/ (2020).

11. Data, software and materials management and sharing policy. *Wellcome Trust* https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy (2017).

12. Open Research. *UK Research and Innovation website* https://www.ukri.org/our-work/supporting-healthy-research-and-innovation-culture/open-research/ (2020).

13. Kuhn, S. & Schlörer, N. E. 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.* **53**, 582–589 (2015).

14. Steinbeck, C., Krause, S. & Kuhn, S. NMRShiftDB-constructing a free chemical information system with open-source components. *J. Chem. Inf. Comput. Sci.* **43**, 1733–1739 (2003).

15. Robien, W. Computer-assisted peer reviewing of spectral data: the CSEARCH protocol. *Monatshefte für Chemie - Chemical Monthly* **150**, 927–932 (2019).

16. Elyashberg, M. & Argyropoulos, D. Computer Assisted Structure Elucidation (CASE):

Current and future perspectives. *Magn. Reson. Chem.* (2020) doi:10.1002/mrc.5115.

17. Han, Y. & Steinbeck, C. Evolutionary-algorithm-based strategy for computer-assisted structure elucidation. *J. Chem. Inf. Comput. Sci.* **44**, 489–498 (2004).
18. Nuzillard, J.-M. & Massiot, G. Computer-aided spectral assignment in nuclear magnetic resonance spectroscopy. *Anal. Chim. Acta* **242**, 37–41 (1991).
19. Jansson, P.-E., Stenutz, R. & Widmalm, G. Sequence determination of oligosaccharides and regular polysaccharides using NMR spectroscopy and a novel Web-based version of the computer program CASPER. *Carbohydr. Res.* **341**, 1003–1010 (2006).
20. Lundborg, M. & Widmalm, G. Structural analysis of glycans by NMR chemical shift prediction. *Anal. Chem.* **83**, 1514–1517 (2011).
21. Castillo, A. M., Bernal, A., Dieden, R., Patiny, L. & Wist, J. 'Ask Ernö': a self-learning tool for assignment and prediction of nuclear magnetic resonance spectra. *J. Cheminform.* **8**, 26 (2016).
22. Cobas, C. NMR signal processing, prediction, and structure verification with machine learning techniques. *Magn. Reson. Chem.* **58**, 512–519 (2020).
23. Chen, D., Wang, Z., Guo, D., Orekhov, V. & Qu, X. Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy. *Chemistry* **26**, 10391–10401 (2020).
24. Sarotti, A. M. Successful combination of computationally inexpensive GIAO 13C NMR calculations and artificial neural network pattern recognition: a new strategy for simple and rapid detection of structural misassignments. *Org. Biomol. Chem.* **11**, 4847–4859 (2013).
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.* **80**, 9371–9378 (2015).

26. Meiler, J., Meusinger, R. & Will, M. Fast determination of 13C NMR chemical shifts using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **40**, 1169–1176 (2000).

27. Binev, Y., Marques, M. M. B. & Aires-de-Sousa, J. Prediction of 1H NMR coupling constants with associative neural networks trained for chemical shifts. *J. Chem. Inf. Model.* **47**, 2089–2097 (2007).

28. Binev, Y., Corvo, M. & Aires-de-Sousa, J. The impact of available experimental data on the prediction of 1H NMR chemical shifts by neural networks. *J. Chem. Inf. Comput. Sci.* **44**, 946–949 (2004).

29. Binev, Y. & Aires-de-Sousa, J. Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *J. Chem. Inf. Comput. Sci.* **44**, 940–945 (2004).

30. Aires-de-Sousa, J., Hemmer, M. C. & Gasteiger, J. Prediction of 1H NMR chemical shifts using neural networks. *Anal. Chem.* **74**, 80–90 (2002).

31. Jonas, E. & Kuhn, S. Rapid prediction of NMR spectral properties with quantified uncertainty. *J. Cheminform.* **11**, 50 (2019).

32. Kang, S., Kwon, Y., Lee, D. & Choi, Y.-S. Predictive Modeling of NMR Chemical Shifts without Using Atomic-Level Annotations. *J. Chem. Inf. Model.* **60**, 3765–3769 (2020).

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

34. Blinov, K. A. *et al.* Performance validation of neural network based (13)c NMR prediction using a publicly available data source. *J. Chem. Inf. Model.* **48**, 550–555 (2008).

35. Unzueta, P. A., Greenwell, C. S. & Beran, G. J. O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning. *J. Chem. Theory Comput.* **17**, 826–840 (2021).

36. Altaf, A. A., Kausar, S. & Badshah, A. Spectral Calculations with DFT. in *Density Functional Calculations* (ed. Yang, G.) (IntechOpen, 2018). doi:10.5772/intechopen.71080.

37. Selzer, P., Gasteiger, J., Thomas, H. & Salzer, R. Rapid access to infrared reference spectra of arbitrary organic compounds: scope and limitations of an approach to the simulation of infrared spectra by neural networks. *Chemistry* **6**, 920–927 (2000).

38. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).

39. Yuanyuan, C. & Zhibin, W. Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks. *Chemometrics Intellig. Lab. Syst.* **181**, 1–10 (2018).

40. Zhang, X. *et al.* Understanding the learning mechanism of convolutional neural networks in spectral analysis. *Anal. Chim. Acta* **1119**, 41–51 (2020).

41. Chen, M., Chen, S. & Zhu, Y. Molecular Structure Prediction Using Infrared Spectra Category : Physical Sciences , CS 229 : Fall 2017. (2017).

42. Patiny, L. ChemInfo. *ChemInfo* http://www.cheminfo.org/Spectra/IR/Exercises/Browse_Spectra/index.html.

43. Selzer, P. Infrared data correlations with chemical structure. *Encyclopedia of Computational Chemistry* (2002) doi:10.1002/0470845015.cca054.

44. IR simulator. *IR simulator* https://www2.chemie.uni-erlangen.de/services/telespec/simuframe/index.html.

45. Selzer, P. IR spectra simulation and information processing on the WWW. *Chimia* **52**, 678–682 (1998).

46. Hemmer, M. C. IR-Spektren: Simulation und Datenbank im Internet. *Nachr. Chem.* **48**,

950–953 (2000).

47. Szantay, C. & Jr. *Anthropic Awareness: The Human Aspects of Scientific Thinking in NMR Spectroscopy and Mass Spectrometry*. (Elsevier Science, 2015).

48. Tremouilhac, P. *et al.* Chemotion ELN: an Open Source electronic lab notebook for chemists in academia. *J. Cheminform.* **9**, 54 (2017).

49. Tremouilhac, P. *et al.* The Repository Chemotion: Infrastructure for Sustainable Research in Chemistry*. *Angew. Chem. Int. Ed Engl.* **59**, 22771–22778 (2020).

50. Tremouilhac, P., Huang, P. C. & Lin, C. L. Chemotion repository, a curated repository for reaction information and analytical data. *Chemistry - Methods* **1**, 8–11 (2021).

51. Huang, Y.-C., Tremouilhac, P., Nguyen, A., Jung, N. & Bräse, S. ChemSpectra: A Web-based Spectra Editor for Analytical data. *Research Square* (2020) doi:10.21203/rs.3.rs-44215/v1.

52. Helmus, J. J. & Jaroniec, C. P. Nmrglue: an open source Python package for the analysis of multidimensional NMR data. *J. Biomol. NMR* **55**, 355–367 (2013).

53. Landrum, G. RDKit, https://www.rdkit.org, Date accessed: May 22, 2023. https://www.rdkit.org.

54. Ertl, P. An algorithm to identify functional groups in organic molecules. *J. Cheminform.* **9**, 36 (2017).

55. Chemotion ELN contributors. *Chemotion Electronic Lab Notebook (ELN)*. (Zenodo, 2024). doi:10.5281/ZENODO.1054134.

56. Huang, P.-C. *et al. Chemotion Repository*. (Zenodo, 2024). doi:10.5281/ZENODO.3755759.

57. Huang, J. *chem-dl-ir: Function group predictions from an IR spectrum;*

https://github.com/ComPlat/chem-spectra-app. (Github).

58. Jason Huang, P. C. H. *ComPlat/chem-dl-ir: chem-dl-ir v1.0.0*. doi:10.5281/zenodo.10654755.