Multimodal Techniques for Detecting Alien Life using Assembly Theory and Spectroscopy

Michael Jirasek†1, Abhishek Sharma†1, Jessica R. Bame†1, Nicola Bell[1], Stuart M. Marshall,[1] Cole Mathis[1], Alasdair Macleod,[1] Geoffrey J. T. Cooper[!], Marcel Swart[2,3], Rosa Mollfulleda[2], Leroy Cronin*[1]

†Equal contribution; *Corresponding author *Lee.Cronin@glasgow.ac.uk*

[1] School of Chemistry, The University of Glasgow, University Avenue, Glasgow G12 8QQ, UK.

[2] University of Girona,Campus Montilivi (Ciencies), c/M.A. Capmany 69, 17003 Girona Spain

[3]ICREA, Pg. Lluis Companys 23, 08010 Barcelona, Spain

Detecting alien life is a difficult task because it's hard to find signs of life that could apply to any life form. However, complex molecules could be a promising indicator of life and evolution. Currently, it's not possible to experimentally determine how complex a molecule is and how that correlates with information-theoretic approaches that estimate molecular complexity. Assembly Theory has been developed to quantify the complexity of a molecule by finding the shortest path to construct the molecule from simple parts, revealing its molecular assembly index (MA). In this study, we present an approach to rapidly and exhaustively calculate molecular assembly and explore the MA of over 10,000 molecules. We demonstrate that molecular complexity (MA) can be experimentally measured using three independent techniques: nuclear magnetic resonance (NMR), tandem mass spectrometry (MS), and infrared spectroscopy (IR), and these give consistent results with good correlations. By identifying and counting the number of absorbances in IR spectra, carbon resonances in NMR, or molecular fragments in tandem MS, the molecular assembly index of an unknown molecule can be reliably estimated from experimental data. This represents the first experimentally quantifiable approach to defining molecular assembly, a reliable metric for complexity, as an intrinsic property of all molecules and can also be performed on complex mixtures. This paves the way to use spectroscopic techniques to unambiguously detect alien life in the solar system, and beyond on exoplanets.

The exploration of chemical space reveals the striking fact that most molecules greater than molecular weight of 300 Da, which are not simple oligomers, are all connected to the existence of life on earth.[1] This is because complex molecules such as natural products[2] are too complex to form by chance in any detectable abundance, and therefore can only be made by the complex biochemical pathways found in biological cells. Currently, the exploration of complex chemical space is done *in-silico*[3,4] and this focusses on chemical structure,[5] topological features,[6] application-specific physicochemical descriptors and graph theory and tends to explore medicinal chemical space for drug discovery and development.[7] In this regard pharmaceuticals can also be considered to be biosignatures, or more specifically technosignatures, since many are complex and would not have been made without humans using technology.[8–10] In addition to target selectivity, synthetic accessibility is important to explore the complexity of the molecule.[11] There are many competing notions of chemical complexity[12], which have led to different algorithmic methodologies being developed using metrics based on molecular weight, counting chiral centres or primarily focusing on substructure properties etc.[13–15] However, with the recent development in automated chemical synthesis,[16] a proxy for complexity is required that is fast to estimate molecular complexity directly from the acquired experimental data, instead of performing complete structure elucidation. Additionally, for biosignature detection,[17] it is important that the complexity metric can be estimated directly from the experimental data without any assumptions about the local environment or chemistry due to the minimalistic information available for an unknown sample.

Recently, we developed a novel approach to quantify and explore the complexity of molecules using Assembly Theory (AT).[18] Assembly Theory estimates the complexity of a molecule by quantifying the minimum constraints required to construct an object from the building blocks. The assembly pathway gives the shortest path to create an object in the absence of physical constraints and reusing the substructures formed along the pathway. The complexity of an object is therefore defined by the number of steps along the assembly pathway and is called the Assembly Index,[19] which for molecules is called

Molecular Assembly (MA). Among the other complexity measures, MA is unique as it has been shown that it can be determined experimentally *via* tandem mass spectrometry.[20] To date, all other approaches to measure molecular complexity cannot be estimated from experimental measurements, and instead require the formula and connectivity of the molecule to be known.[21] The Molecular Assembly (MA)[22,23] for a molecule is computed by representing the molecule as a graph and performing an algorithmic search to find the shortest pathway to construct the graph by reusing previously made structures along the pathway, see **Fig. 1A**. Thus, various constraints in the molecular graph are found along the pathway to quantify the complexity of the molecule. This means that if you take a given target molecule with *n* bonds, and you take *n* copies of the molecule and break one different bond in each of the *n* copies, it is possible to deduce the molecular assembly index for that molecule if you remove identical units thereby only counting the number of unique parts, see **Fig. 1**.

In previous work, we used Tandem Mass Spectrometry (MS/MS) for the experimental measurement of MA and were able to rank molecules in order of their complexity by placing them on a scale where molecules beyond a MA of 15 were considered to be biosignatures for life-detection. Experimentally, over a range of high MA molecules, it was demonstrated that a high correlation exists between MS^2 peaks and computed MA values[20]. However, a key limitation of the technique is its requirement for the molecules to be in a charged state in the gas phase, which limits the search space of complex molecules.[24] Herein, we developed experimental measurement strategies to infer molecular complexity using MA by using IR and NMR spectroscopies. Using both simulated and experimental data, we demonstrate that MA can be experimentally inferred over a wide range of complex molecules as well as mixtures. Additionally, we demonstrate that by combining multiple spectroscopic techniques into one measure, the MA prediction can be improved further.

Infrared Spectroscopy (IR) is routinely used to confirm the presence of specific bond types in molecules by observing their characteristic vibrational energies in higher energy ranges (1500–3600 cm^{-1}).

Vibrational motion corresponding to those absorption bands is typically of local nature, for example, a stretch vibration of one bond. Contrary, lower energy (fingerprint) region 400–1500 cm^{-1} typically possesses a plethora of absorption bands, without direct easy interpretation toward structure elucidation.[25] These modes include various collective motions, bending vibrations, and coupled modes. Since the number of different substructures increases with the molecular complexity, we hypothesise that number of unique absorption bands in the fingerprint region can be used to infer the complexity of organic molecules. Moreover, IR has previously been used to fingerprint complex molecular ensembles in their native natural environment.[26]

Nuclear Magnetic Resonance (NMR) spectroscopy provides resonance frequencies of magnetically inequivalent atoms nuclei in the structure. The exact chemical shift of each nucleus of the same element depends on the effective magnetic field experienced by it, strongly influenced by its chemical microenvironments (affected by e.g. bond correlation *via* scalar coupling, or through space (de)shielding effects).[27] NMR has been used in the past to analyse chemical space for fragment screening in drug discovery[28–30] and characterising structural complexity of compound classes[31]. In contrast to mass spectrometry, NMR has minimal solvent limitation allowing to keep the sample in its native state/solvent if desired. NMR is uniquely equipped to address the structural diversity as symmetric (magnetically equivalent) units in an isotropic environment (e.g., in a homogenous solution) possess the same chemical shift (thus not creating duplicated resonances). Further, from the perspective of molecular assembly, the effect of symmetry and bond rotation on NMR spectra was hypothesised to provide near equivalent resonances for duplicated fragments, even if not magnetically equivalent as a result of very similar chemical microenvironment experiences by the fragment. This represents the fact that assembled fragments may be utilised in multiple symmetric positions in a structure without having to 'rebuild' them each time. Therefore, we hypothesise that number of observed NMR resonances will reflect a degree of structural complexity.

Both IR and NMR will agnostically indicate the complexity of a molecule defined by MA since Assembly Theory states that the MA utilizes unique irreducible motifs to construct the molecule that are indicated by the observed spectral features. This suggests that spectroscopy techniques that can quantify the properties of unique environments, and molecular substructures should in principle produce a good correlation with MA. Thus, we hypothesise that with more unique bond types and atomic environments for a given molecule, the larger the number of peaks that should be found in IR and NMR spectra for that molecule, see **Fig. 1**.

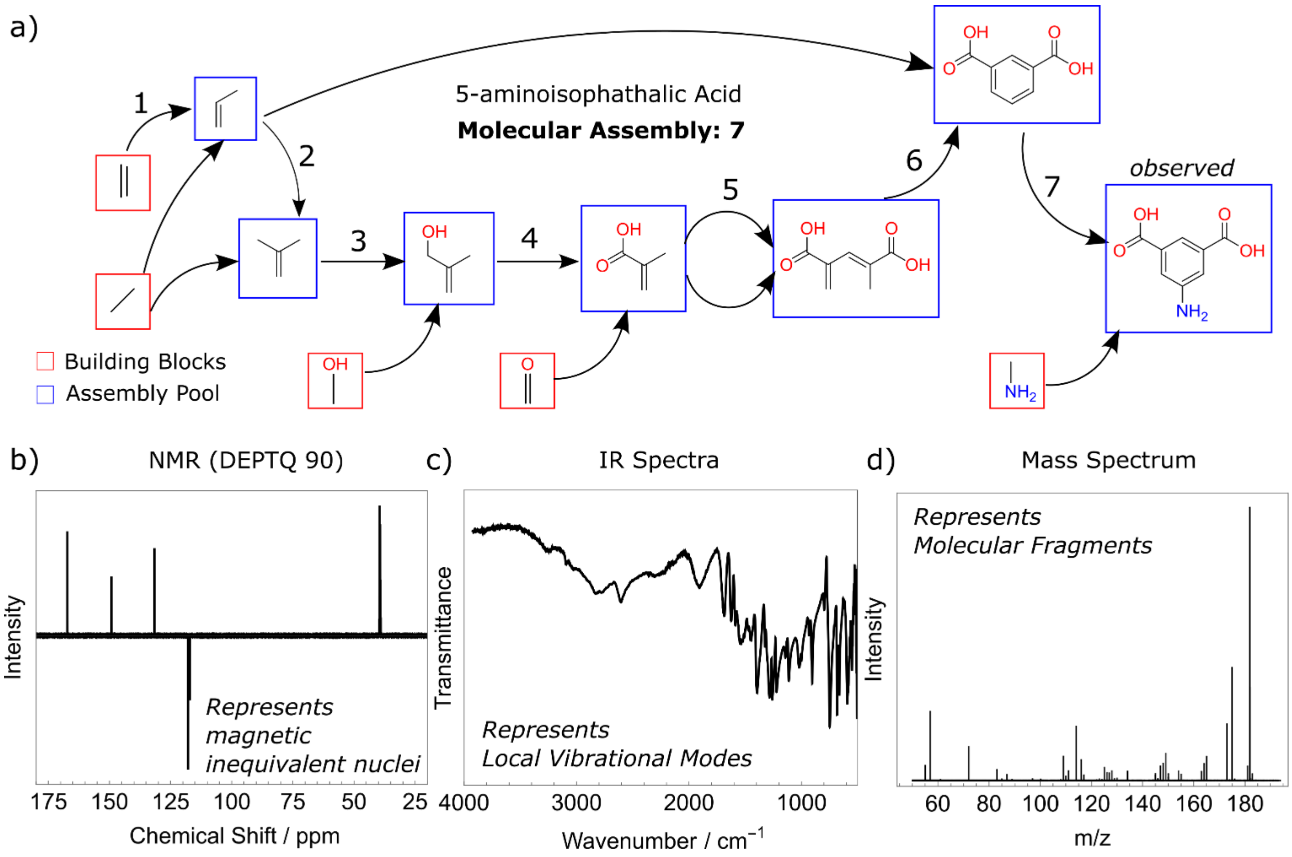


Fig 1. Molecular Assembly of 5-aminoisophathalic acid. (A) Molecular Assembly pathway of 5-aminoisophathalic with a total of 7 steps. The various chemical bonds are considered as fundamental building blocks (shown in red) and the substructures (shown in blue) along the pathways constitute the assembly pool. (B-D) Experimental NMR, IR, and MS^2 spectra of 5-aminoisophathalic acid highlight different features of the molecule from which the molecular constraints and the MA can be inferred.

Calculating Assembly Index from a molecular graph

The Assembly index, and associated minimal assembly pathways, are calculated using an algorithm written in the Go programming language. In prior work[20], the assembly index was calculated using a serial algorithm written in C++ and yielded the "split-branch" assembly index, an approximation that provides a reasonably tight upper bound for the assembly index. The Go algorithm used in this work is a faster algorithm that incorporates concurrency and can provide the exact assembly index if it can be calculated in a reasonable time. The process can also be terminated early to provide the lowest assembly index found so far, which has been found to be a good approximation for the assembly index in most cases.

The assembly index is calculated by iterating over subgraphs within a molecular graph and finding duplicates of that subgraph within the remainder of the molecule. For each of the matching subgraphs found an assembly pathway can be represented by a duplicate structure and a remnant structure (see **Fig. 2**). The remnant structure comprises the original structure with one duplicate removed, and the other "broken off", which ensures that all structures on an assembly pathway that are duplicated will be first constructed. The process can then be repeated recursively with the remnant structure as an input, which may result in more pathways containing two duplicate structures and a smaller remnant. Thus, each pathway is represented by a sequence of duplicated structures and a remnant structure. In this regard it is important to note that molecular assembly uses bonds as building blocks and not atoms. In order to determine the assembly index, we consider that a molecular graph with N bonds could be constructed in $N-1$ steps by adding one bond at a time (the naive MA, or MA_{naive}). Each duplicate structure of size N_{dup} allows us to add that structure in one step, reducing the number of steps compared to MA_{naive} by $N_{dup}-1$. Thus, the MA for a particular pathway is $MA_{naive}-\sum_{dup}(N_{dup}-1)$. For more details, see **SI Section 1**.

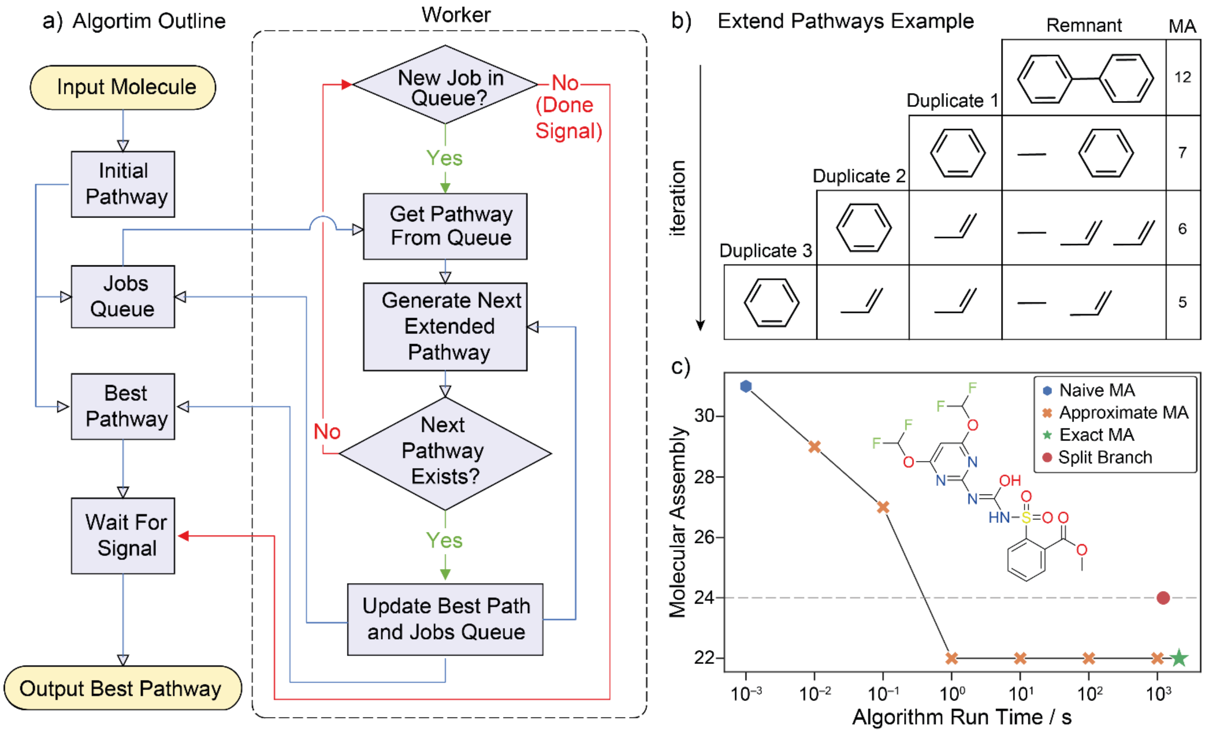


Fig 2. (a) The general structure of the Go assembly algorithm, with a pool of worker extending pathways. Some features are omitted for brevity, such as branch and bound methods to improve efficiency. (b) A sequence of assembly pathways as processed by the Go algorithm. The top pathway is the starting pathway for the molecule shown, and each subsequent pathway is extended from the pathway above. Pathways are generally extended in multiple ways, and only one such sequence of extensions is shown here. (c) An example of MA values found over time for Primisulfuron-methyl, run to completion, and approximated by stopping early at various stages prior. The new algorithm found pathways at the correct MA of 22 by 10 s, significantly before completion at ~2064 s. The red circle shows split branch algorithm performance on the same molecule. The naïve MA (blue hexagon) is calculated trivially for pathways in which one bond is added at a time (placed illustratively at 10^{-3} s, as 0 s cannot be represented on the logarithmic scale).

Inferring Assembly Index using Infrared Spectroscopy

As a first step, we computationally explore the potential for inferring the assembly index from IR absorption. A set of 10,000 molecules were chosen uniformly from the previous dataset[20] of approx. 10^6 molecules with MA. The new algorithm vastly speeded up the calculation and we were able to sample chemical space calculating the MA (previously called Pathway Assembly, calculated using a split-branch algorithm). This was done so that we calculated MA for *ca.* 650 molecules at each MA unit between 2 and 23 MA for each molecule with the new implementation.[32] We calculated the IR spectra of the molecules using an extended semiempirical Tight Binding model implemented in xTB software including geometry optimization and calculating frequency resonances (see **SI Section 2.3**). For each spectrum, we estimated the total number of peaks in the fingerprint region (400–1500 cm^{-1}) assuming a resolution of 2 cm^{-1}. The number of peaks correlated significantly (Pearson correlation coefficient of 0.86) with the calculated MA, yielding a simple prediction function that was phenomenologically derived via linear regression: $\mathrm{MA} = 0.21 \times n_{peaks} - 0.15$ (**Eq. 1**). This observation corroborated our hypothesis that number of absorption peaks in the IR fingerprint region reflects molecular complexity, see **Fig. 3**.

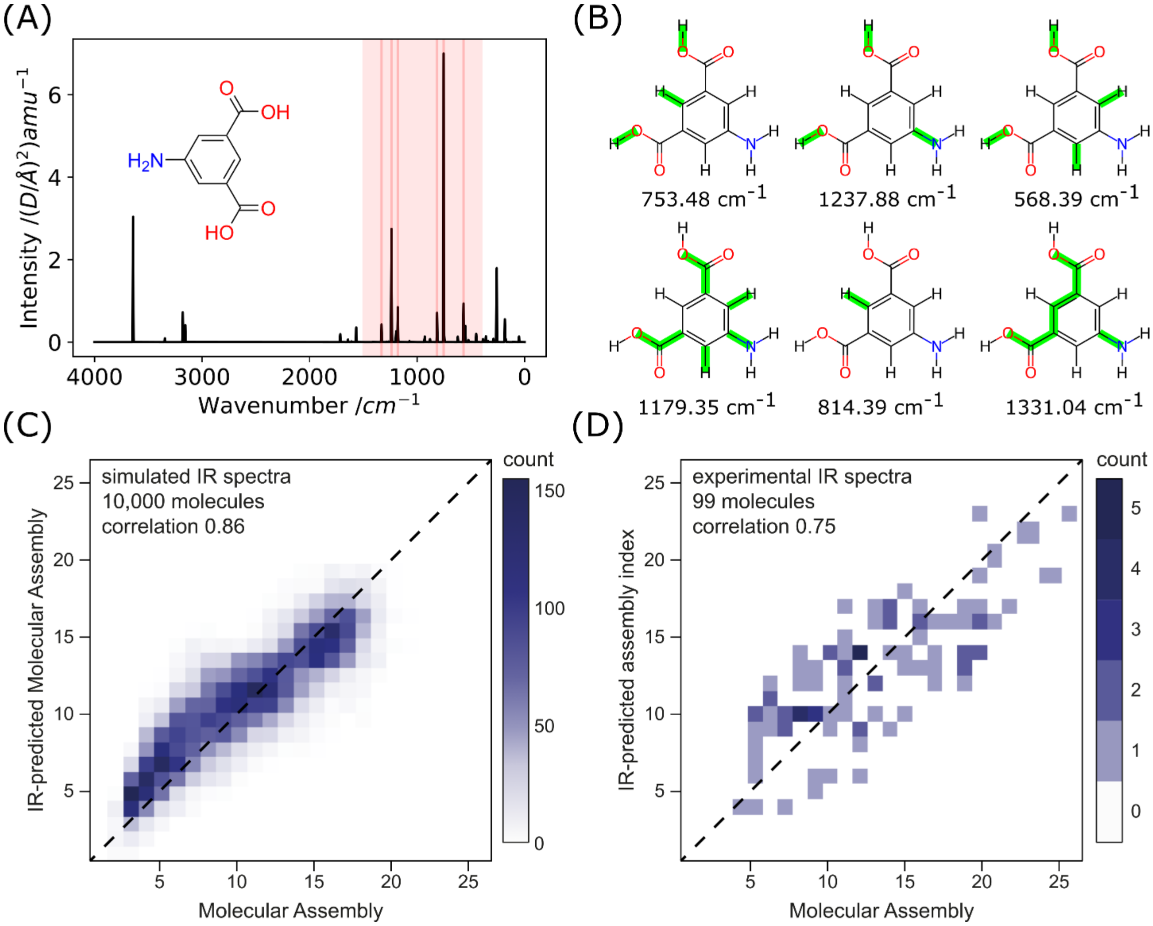


Fig. 3. Inferring Molecular Assembly from infrared spectroscopy. (A) XTB-calculated IR spectrum of 5-aminoisophthalic acid with highlighted fingerprint region (400–1500 cm^{-1}). (**B)** Example of the six most intense vibrational bands in the fingerprint region demonstrating its collective-motion nature. (**C)** Assembly index vs. IR-inferred assembly index estimated from the number of IR peaks in the fingerprint region (400–1500 cm^{-1}) based on XTB calculation on 10,000 molecules (see **Eq. 1**). Correlation between the predicted and expected assembly index is 0.86. **(D)** Assembly index vs. IR-inferred assembly index estimated from the number of IR peaks in the fingerprint region (400–1500 cm^{-1}) based on the experimental measurement on 99 molecules (see **Eq. 2**). Correlation between the predicted and expected assembly index is 0.75.

Further, we expanded the study with experimental validation, using a set of 99 compounds MA over the range 4–26. The experiments were performed using diamond-attenuated total reflectance IR spectroscopy with a resolution of 2 cm^{-1}. The obtained spectra were processed at 50% sensitivity and up to 80% transmittance threshold for selecting peaks using OMNIC software as the coarse filter against low-intensity noise in real spectra. The total number of IR peaks in the fingerprint region (400–1500 cm^{-1}) correlated well with the compounds MA with a 0.75 correlation coefficient. This provided a handle for inferring an assembly index from an experimental IR using a simple linear function: $\mathrm{MA} = 0.45 \times n_{peaks} - 2.3$ (**Eq. 2**). For more details see **SI Section 3**.

Inferring assembly index from NMR spectra

Most organic molecules (by definition) are composed of mainly carbon and hydrogen atoms, we hypothesised that ^{13}C NMR is a practical technique to infer the molecular assembly of organic molecules. This was because the computation of molecular assembly is based upon bonds as building blocks and that NMR will be uniquely able to explore the connectivity within complex organic molecules by exploring and quantifying the types of carbon atom present such as CH_3, CH_2, CH , and C, along with their relative connectivity's. For the experimental measurement, a spectral width within which typically observed ^{13}C nuclei resonances is relatively broad (~200 ppm) was considered, and it is reasonable to assume that inequivalent nuclei of sufficiently different microenvironments would rarely possess the same resonance frequency within a resolution of 0.5 ppm. Further, we expect that magnetically non-equivalent, yet structurally very similar sub-units with the exact environment in nuclei vicinity will be found within the resolution width (see **SI Section 2.2**). Observing such overlap will reflect the unit's similarity and the peaks will not be overcounted as the corresponding substructures likely share the assembly space (the space of motifs that are used to construct the target) and do not contribute to the assembly index (e.g. repeating units of the polymer chain such as –CH_2–).

Further information that can be experimentally extracted from the ^{13}C-NMR spectrum is the classification of the carbon nuclei by the number of attached hydrogens. Based on the assembly theory, we hypothesise that the presence of carbons with no attached hydrogens (for clarity will be referred to as *quaternary*) reports most significantly on the molecular complexity as such centres are highly connected to four different atoms, but also can be connected to a range of different heteroatoms. Thus, these centres are hard to produce, and needed a large number of constraints to construct them. Analogously, we hypothesise that the more hydrogens are attached to the carbon, the less localized information it stores and hence, contributes less to the molecular assembly. From the experimental point of view, the classification of carbon nuclei by the number of attached hydrogens can be experimentally achieved using standard DEPTQ-90 and -135 routines, which provide information about the number of hydrogens attached to the carbon *via* the hydrogen-carbon coupling.[33,34]

Theoretical investigation

To test our hypothesis, we examined a set of predicted ^{13}C NMR spectra of 10,000 molecules (the same set as in the case of theoretical IR investigation). We have used the established predicting tool NMRShiftDB employing the Hierarchical Organization of Spherical Environments (HOSE) method.[35] An example of NMR prediction for two molecules (5-aminoisophathalic acid ($\mathrm{MA} = 7$) and Quinine ($\mathrm{MA} = 16$) with various carbon atoms labelled is shown in **Fig. 4(A&B)**. We classified the carbons by the number of hydrogens attached to them and summed the number of predicted peaks of a certain type assuming a bin width of 0.5 ppm. We performed multivariate linear regression (weighing out differently different types of carbons) and provided a model with a good correlation of 0.87, see **Fig. 4C.** The formula for inferring the assembly index from the number of found peaks associated with individual carbon types was phenomenologically derived via linear regression to be: $\mathrm{MA} = 1.3 \times \mathrm{C} + 0.8 \times \mathrm{CH} + 0.6 \times \mathrm{CH_2} + 0.3 \times \mathrm{CH_3} + 2.1$ (**Eq. 3**), where C (quaternary), CH (tertiary), CH_2 (secondary) and CH_3 (primary) are the number of binned (by 0.5 ppm) ^{13}C resonances of carbons with none, one, two or three hydrogens attached, respectively. This observation on a large dataset significantly corroborates our prediction that quaternary carbons possess the highest degree of constraints and have the highest potential to report on molecular complexity.

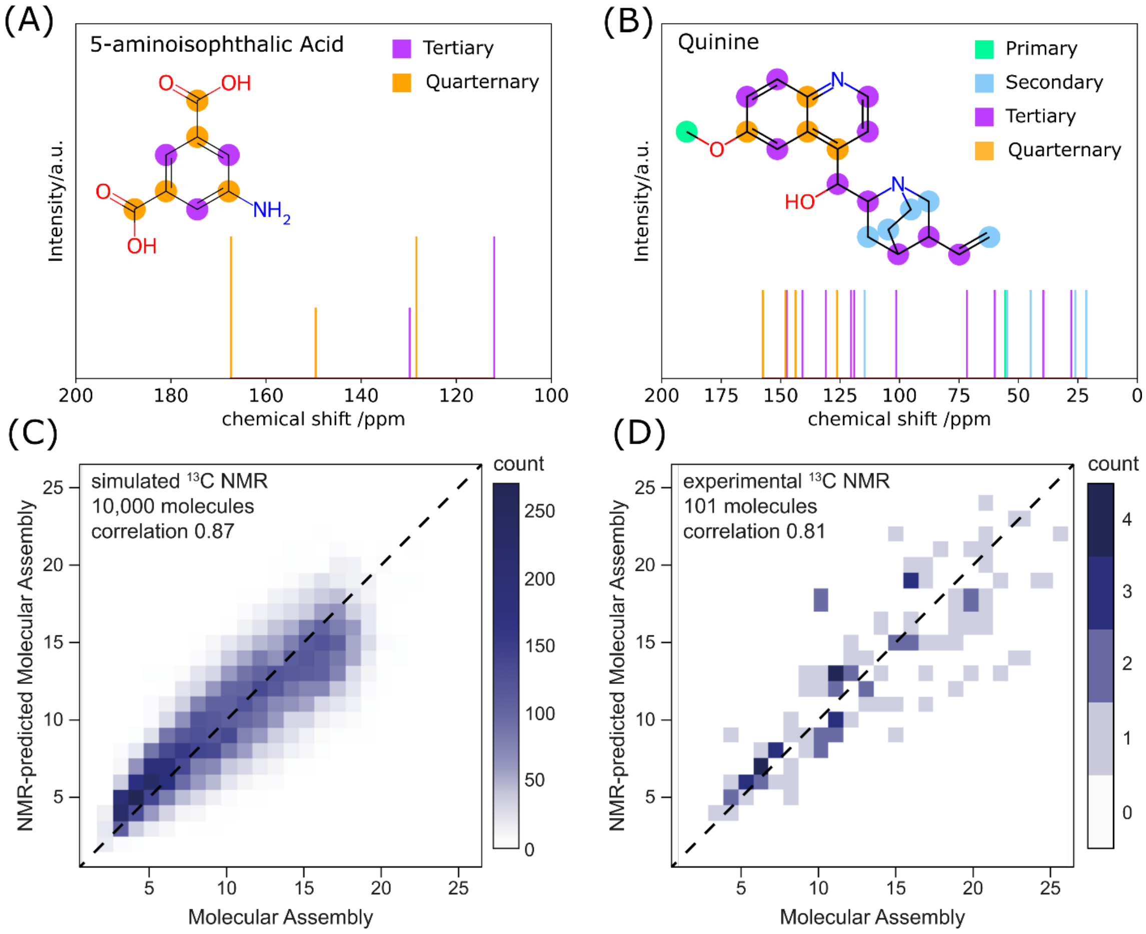


Fig. 4. Inferring molecular complexity from ^{13}C NMR spectra. **(A)** and **(B)** shows the predicted ^{13}C NMR spectrum of 5-aminoisophathalic acid and quinine, with highlighted different types of carbons. **(C)** Assembly index vs. NMR-inferred assembly index estimated from the number of different types of carbons (see **Eq. 3**) based on NMRshiftDB calculation on 10,000 molecules. The correlation between the predicted and expected assembly index is 0.87. **(D)** Assembly index *vs.* NMR-inferred assembly index estimated from the number of different types of carbons experimentally on 101 molecules, using the same model as in the theoretical set. The correlation between the predicted and expected assembly index is 0.81.

Experimental validation

For experimental validation, we have assessed 101 compounds covering a range of assembly index of 3–26. We have acquired ^{13}C NMR spectra and experimentally assigned the carbon type (C, CH, CH_2 and

CH_3) *via* DEPTQ-90 and DEPTQ-135. The correct assignment was further cross-validated with ^1H-^{13}C HSQC since occasionally post processing of DEPTQ spectra can result in the inversion of the peaks phase. As the number of peaks is a simple and reliable measure directly comparable to the experimental observable property, we could test the trained model (**Eq. 3**) directly on independently chosen experimental molecules. Testing the trained model provided a good correlation of 0.81, see **Fig. 4D**. Allowing the change in the multivariate regression on the experimental set could provide an even better correlation of 0.86 (**see SI Section 4**), however, we have considered using the model train on a large dataset as the more robust model, less biased by the sampling of the chemical space.

Measuring the complexity of mixtures

Measuring the complexity of mixtures is essential for analyzing samples of unknown origins for biosignatures, exploring natural products, and closed-loop robotic platforms performing open-ended exploration without setting any explicit target. Additionally, the information content of a mixture requires complexity and the relative abundance of individual compounds present in the mixture[19]. Experimentally, estimating the complexity of an unknown mixture is a challenge for analytical techniques as the predicted complexity could either be a function of a large number of compounds of lower complexity or a smaller number of compounds with large complexity.[13,36,37] Here, we demonstrate two different techniques for analyzing mixture complexity utilizing ^{13}C DOSY spectroscopy and LC-MS/MS based on previous work.

Analysing Mixtures using NMR

The NMR technique is well-equipped for analysing complex mixtures.[26] To experimentally deconvolute the mixture of chemical resonances to their individual components, we investigated ^{13}C DOSY spectroscopy on mixtures, separating individual compounds *via* their diffusion coefficient.[38] Together with the experimental assignment of the carbon types, we could predict the assembly index of each component in the mixture, using the same logic as for the individual compounds. An example of such workflow is assessing a mixture of 5-aminoisophathalic acid and quinine: Using the DEPTQ routines, different types of carbons were assigned (**Fig. 5A**). Using ^{13}C DOSY, the peaks were assigned to two different individual components (**Fig. 5B**). Finally, the assembly index of the species based on the number of carbon resonances was inferred to be 8 and 19, in reasonably good agreement with the expected real value of 7 and 16, respectively (**Fig. 5C**). For more details, see **SI Section 6.1**.

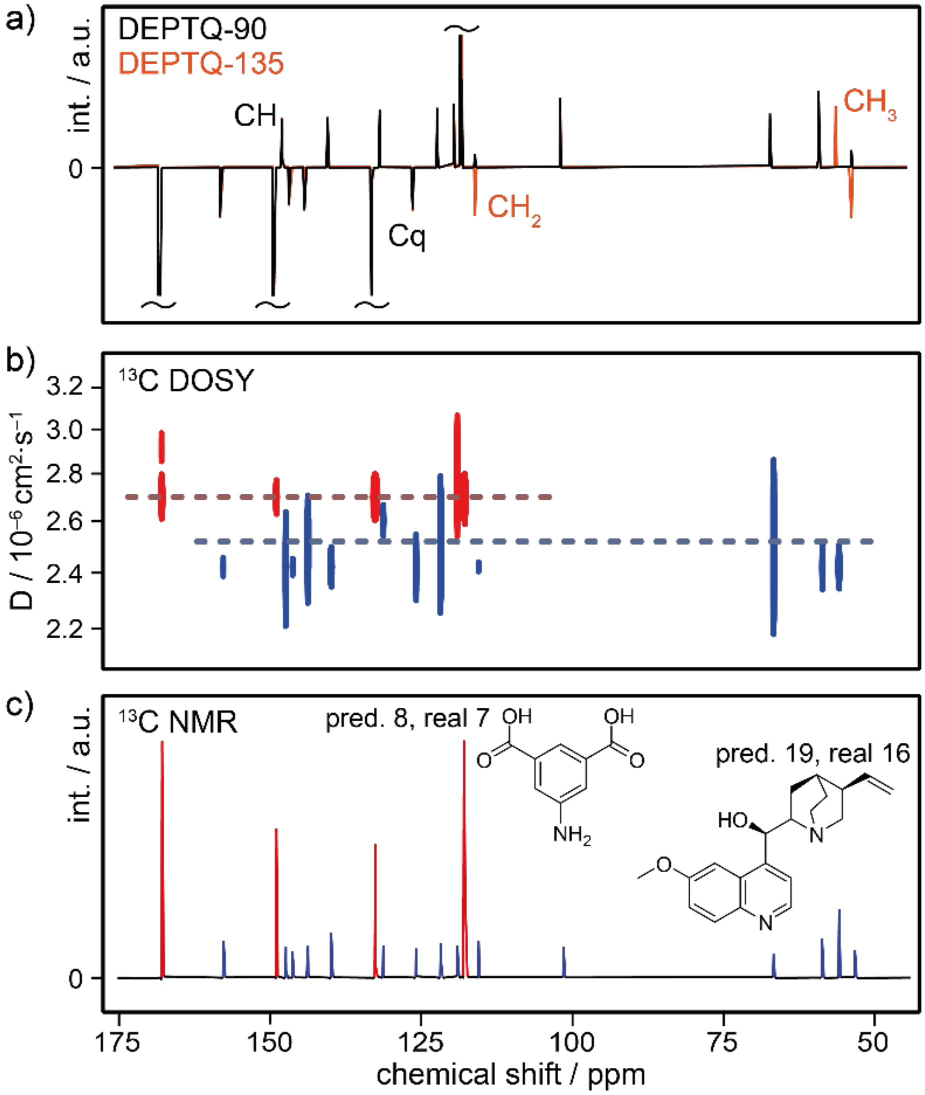


Fig. 5. Example of deconvolution of mixture complexity. (A) Overlay of spectra using DEPTQ-90 and DEPTQ-135 methods to differentiate different types of carbons. (B) ^{13}C-DOSY deconvolution of peaks to the individual compound assignment *via* diffusion. (C) Assigned deconvolution of ^{13}C-NMR spectra, with peaks for 5-aminoisophathalic acid in red and quinine in blue.

Analysing Mixtures using Tandem Mass Spectrometry

It has been shown that tandem mass spectrometry can be used to assess the complexity of molecules with a high degree of certainty.[20] Building on the prior work, we used the new MA algorithm[32] to calculate the MA values for correlation with the experimental data from the available dataset, see **Fig. 6**.

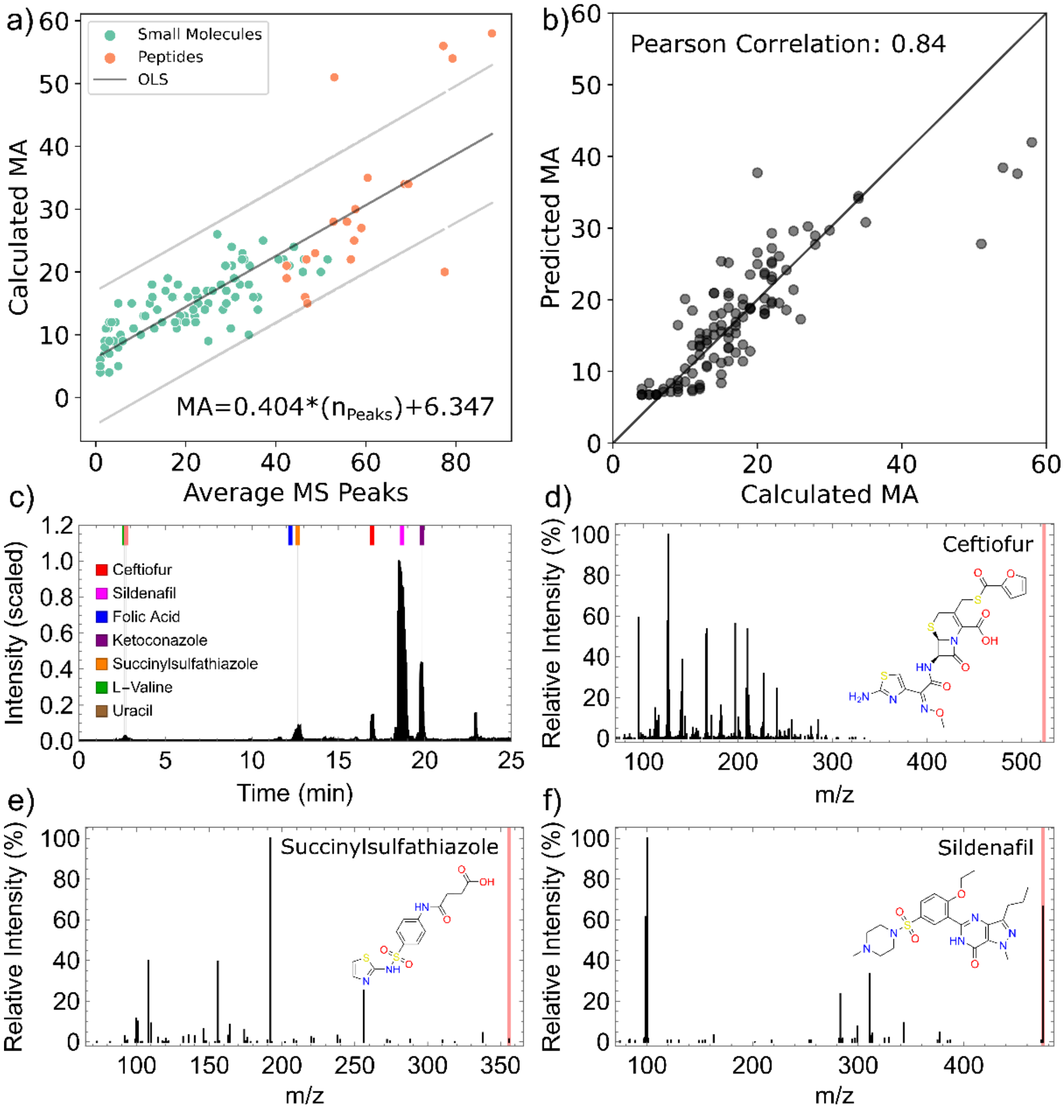


Fig. 6 LC-MS Mixture Analysis allows the determination of MA number for individual molecules. (A) Ordinary Linear Regression between calculated MA and average MS^2 peaks (from previous work), including small molecules and peptides. (B) Calculated and predicted MA using the linear regression model on the same dataset with a Pearson Correlation Coefficient of 0.84. (C) Chromatogram of a mixture of ten molecules with seven molecules identified. (D–F) MS^2 spectra of Ceftiofur, Succinylsulfathiazole, and Sildenafil respectively showing compounds from the mixture can be resolved.

We performed ordinary linear regression analysis on this dataset and show a good correlation between the previously estimated number of MS^2 peaks and calculated MA with Pearson's correlation coefficient of 0.84. The relationship between MA and the number of MS^2 peaks can be described as $\mathrm{MA} = 0.4 \times n_{\mathrm{peaks}} + 6.3$ **(Eq. 4)**, see **Fig. 6A** and **6B**. As an extension, to demonstrate its capability to address the complexity of individual compounds in a mixture, we prepared a solution of 10 different compounds with MAs in the range 5–26 including ceftiofur, sildenafil, folic acid, ketoconazole, succinylsulfathiazole, L-valine, and uracil. LC-MS/MS analysis of this mixture was performed which allowed the isolation of individual species based on retention time and an estimation of their MA values by basic peak counting of their respective MS^2 spectra and determining MA using **Eq. 4,** see **SI Section 6.2** for details. **Fig. 5C** shows the chromatogram from LC distinguishing seven molecules and **Fig. 6D-E** shows respective MS2 fragmentation spectra for three individual molecules.

Combining analytical techniques for Molecular Assembly inference

Molecular constraints are probed by different physical interactions depending on the spectroscopic techniques, which independently have been shown to correlate with Molecular Assembly. In general, due to different limitations in the considered spectroscopic techniques (NMR, IR, and MS), individual spectral features of a specimen of unknown origin can be considered biased or *contaminated*. For example, MS/MS fragments distribution is biased by the strengths of the different chemical bonds, which molecular assembly calculation does not consider. Similarly, ^{13}C NMR spectroscopy might not fully reflect the MA should the constraints be realised through heteroatoms; further diastereotopic carbons can be overcounted although considered equivalent. IR fingerprint region can contain overtones of the functional groups, causing the peak overcount. All herein listed examples responsible for variance in the correlation with the MA have principally different physical interactions. We, therefore, hypothesised that a combination of the analytical techniques can increase confidence in the MA inference, see **Fig. 7**.

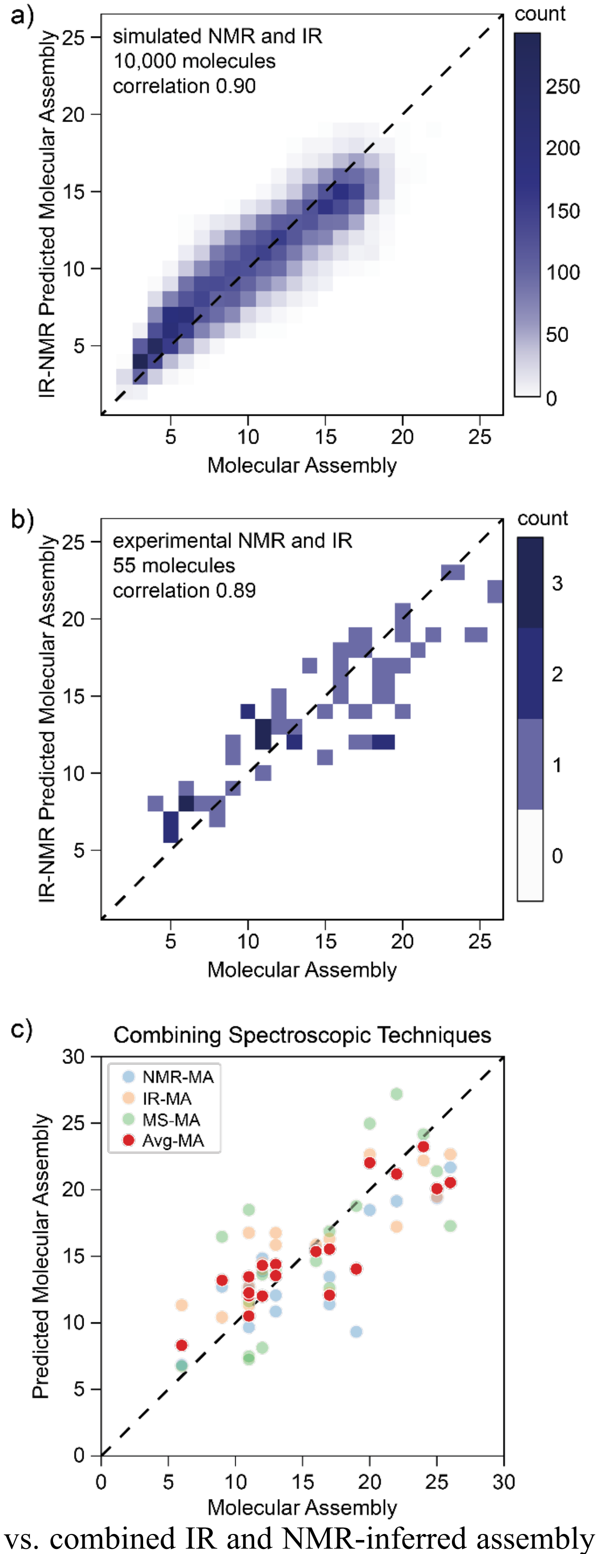


Fig. 7 (A) Assembly index vs. combined IR and NMR-inferred assembly index (using weights of 0.55 and 0.45 from NMR and IR, respectively) based on 10,000 calculated spectra showing an increased correlation of 0.90. **(B)** Assembly index vs. combined IR and NMR-inferred assembly index (using weights of 0.7 and 0.3 from NMR and IR, respectively) based on 55 experimental spectra showing an increased (relative to the individual components) correlation of 0.89. **(C)** Assembly index *vs*. individual and combined IR, NMR and MS-inferred assembly index based on the 19 molecules where all three experimental datasets were available.

On the set of 10,000 calculated NMR and IR spectra, we have examined our hypothesis that combined information can provide a more reliable MA prediction. We have used the same models (**Eq. 1** and **Eq. 3**) for the individual spectroscopic techniques and allowed them to optimise for their relative weighting. The combined model provided a higher correlation of 0.91 using the weighted average of 0.55×NMR and 0.45×IR inferred MA, see **Fig. 7A**. Further, we have validated this approach on the available intersection of the experimental NMR and IR data, comprising 55 molecules. The combined model provided a higher correlation of 0.89 using a combination of 0.7×NMR and 0.3×IR inferred MA, see **Fig. 7B**.

Lastly, we have explored the combination of all three techniques to infer MA. Where the experimental dataset was available, we used predicted MA from all three techniques, both as individual components and as an average and compared it with the expected MA value (**Fig. 7C**). Although the average value might not always provide a better estimate than certain individual components, it provides more robust prediction not susceptible to large deviations coupled to individual physical phenomena. This demonstrates that for inferring MA of unknown species it would be preferable to apply more experimental techniques and acquire an average of their MA predictions.

The work here shows that the general concept of measuring molecular complexity as a function of the number of different parts in a molecule, using spectroscopic measurements, gives a very strong correlation with the theoretical assembly complexity. This is important since it means we can use experimental measurements on environmental samples to read out the amount of selection and evolution that the samples have been subjected to making this approach suitable for the search for new life on earth and life beyond earth.

Conclusion

We have demonstrated on a set of 10,000 simulated and approximately 100 experimental IR and NMR spectra that it is possible to predict the MA of compounds, without their structural elucidation. This is particularly useful for molecules from unknown origins and cases when a fast metric for probing complexity is required. In the case of IR, the constraints and molecular complexity are reflected by the number of peaks in the fingerprint region and their simple summation can be used to predict molecular complexity. In NMR, we have shown that the weighted sum of the number of carbon resonances, sorted by the number of hydrogens attached to them, provides a good prediction of MA. We found that the fewer hydrogens attached to the carbon, the higher weight it possesses for the MA prediction. This finding corroborates our interpretation based on Assembly Theory that the quaternary carbons effectively encode the most information whereas the primary carbons, which have more hydrogen atoms, are least encoded and hence contributes less to the molecular assembly. Finally, we have demonstrated that is possible to address the complexity of the components in mixtures. We have shown that ^{13}C DOSY facilitates deconvolution of the ^{13}C NMR signals to their individual compounds based on diffusion coefficient which MA can be inferred. Similarly, LC-MS/MS can discern the individual components of a mixture and be used to predict MA values. In combination with the previously reported method of using tandem mass spectrometry to measure molecular complexity, NMR and IR provide an important tool to predict and cross-validate experimentally measured MA.

These findings are of particular significance for the development of missions looking for life in our solar system.[39] NASA has already managed to put several mass spectrometers on Mars,[40] and several mass specs have been in the solar system including on the Cassini probe which visited Saturn and Enceladus.[41] Dragonfly is set to visit Titan, launching in 2026 and arriving in 2034, is important since it will be a mobile mass spectrometer that flies around Titan.[42] Of critical importance will be the ability to resolve high molecular weight compounds (300-600 Da) with the possibility of generating in situ fragmentation.

In this instance, the use of Assembly Theory when analysing the data will allow us to put some limits on the complexity of the molecules found on Titan. Further afield, as exo-planet spectroscopy becomes more advanced, it will be possible to look for infra-red signatures associated with exoplanets. Whilst very high-resolution imaging will not be possible, without using gravity lensing (using the sun),[43] it might be possible to build a network of chemistries that could be constructed from the small molecule disequilibria found on any exoplanets.[17] Whilst these applications seem far away, it is only now with assembly theory making firm and experimental predictions about the complexity of molecules that can result from an evolutionary or informational process, that we can seriously contemplate truly agnostic biosignature searches now we have validated the measurement of molecular complexity across three different experimental domains.

Experimental Section:

Infrared Experimental Setup: IR spectra were acquired on a Thermo Scientific Nicolet iS5 with Specac Golden Gate Reflection Diamond ATR System. All data were processed with Thermo Scientific OMNIC 8.3.103. All samples were measured in the native state at room temperature (solid state unless liquid at room temperature).

NMR Experiment Setup: NMR data were acquired on a Bruker Ascend Aeon 600 MHz NMR spectrometer with a DCH cryoprobe (^{13}C + ^{1}H channels) at 300 K unless otherwise stated in which case a Room temperature BBFO probe head (^{1}H + ^{19}F-^{183}W channels) was used. ^{1}H NMR were acquired using 16 scans, spectral width 20 ppm and relaxation delay 2 s. Spectra on the ^{13}C channel were acquired with a spectral width of 200 ppm. The ^{13}C NMR spectra were acquired using 16 scans and a relaxation delay of 0.8 s. The DEPTQ routines were carried out using 16 scans and a relaxation delay of 1 second. The ^{13}C DOSY spectra were acquired using 256 scans, a relaxation delay of 8 seconds and a 500-1000 µs gradient pulse. All spectra were processed using Bruker Topspin 3.6 and Mestrenova 14.1.1. The spectra were phase and baseline corrected and calibrated relative to the residual solvent peak. Residual solvent peaks were not included in resonance counts. Unless otherwise stated samples were prepared in DMSO-d_6 at the concentration stated in the ESI.

Supplementary Data

This describes the algorithm for calculating assembly index for molecules (molecular assembly – MA), the details of the theoretical calculations for NMR, Infra-Red (xTB and DFT simulations), sample preparations for the experimental data collection, experimental IR and NMR data, the regression analysis, mixture analysis. The Molecular Assembly calculator called *AssemblyGo* was written in GO programming language (https://github.com/croningp/assembly_go). The codes used for processing data and further details can be found at https://github.com/croningp/molecular_complexity.

Acknowledgements

The authors gratefully acknowledge financial support from the John Templeton Foundation (Grant 60025), EPSRC (Grant Nos EP/L023652/1, EP/R01308X/1, EP/J015156/1, EP/P00153X/1), the Breakthrough Prize Foundation and NASA (Agnostic Biosignatures award #80NSSC18K1140), MINECO (project CTQ2017-87392-P) and ERC (project 670467 SMART-POM). We would like to thank Sara. I. Walker and Estelle M. Janin (ASU) for comments on the manuscript.

Author Contributions

MJ generated calculated NMR and IR spectra and interpreted all data; AS generated calculated NMR and IR spectra and interpreted all data; JB collected the NMR and IR experimental data and did preliminary fitting with NB; CM helped with the assembly algorithm development and wrote software interpreting the mass spectrometry data; AM acquired the mass spectrometry data for the mixture; GJTC provided some samples for the blinded tests; SMM developed the assembly go program; MS and RM generated IR data from DFT analysis. LC developed assembly theory, conceived the idea, raised the funding, and supervised the research. LC wrote the paper with input from all the authors.

References

1. Mikulak-Klucznik, B. *et al.* Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
2. Pilkington. A Chemometric Analysis of Deep-Sea Natural Products. *Molecules* **24**, 3942 (2019).
3. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
4. Adams, K. & Coley, C. W. Equivariant Shape-Conditioned Generation of 3D Molecules for Ligand-Based Drug Design. (2022) doi:10.48550/ARXIV.2210.04893.
5. Böttcher, T. From Molecules to Life: Quantifying the Complexity of Chemical and Biological Systems in the Universe. *J. Mol. Evol.* **86**, 1–10 (2018).
6. Isert, C., Atz, K. & Schneider, G. Structure-based drug design with geometric deep learning. (2022) doi:10.48550/ARXIV.2210.11250.
7. *Complexity in chemistry: introduction and fundamentals*. (Taylor & Francis, 2003).
8. Clemons, P. A. *et al.* Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci.* **107**, 18787–18792 (2010).
9. Méndez-Lucio, O. & Medina-Franco, J. L. The many roles of molecular complexity in drug discovery. *Drug Discov. Today* **22**, 120–126 (2017).
10. González-Medina, M. *et al.* Chemoinformatic expedition of the chemical space of fungal products. *Future Med. Chem.* **8**, 1399–1412 (2016).
11. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).

12. Sheridan, R. P. *et al.* Modeling a Crowdsourced Definition of Molecular Complexity. *J. Chem. Inf. Model.* **54**, 1604–1616 (2014).

13. Böttcher, T. An Additive Definition of Molecular Complexity. *J. Chem. Inf. Model.* **56**, 462–470 (2016).

14. Barone, R. & Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **41**, 269–272 (2001).

15. F., K. *et al.* Molecular complexity determines the number of olfactory notes and the pleasantness of smells. *Sci. Rep.* **1**, 206 (2011).

16. Jiang, Y. *et al.* An artificial intelligence enabled chemical synthesis robot for exploration and optimization of nanomaterials. *Sci. Adv.* **8**, eabo2626 (2022).

17. Schwieterman, E. W. *et al.* Exoplanet Biosignatures: A Review of Remotely Detectable Signs of Life. *Astrobiology* **18**, 663–708 (2018).

18. Marshall, S. M., Murray, A. R. G. & Cronin, L. A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **375**, 20160342 (2017).

19. Sharma, A. *et al.* Assembly Theory Explains and Quantifies the Emergence of Selection and Evolution. Preprint at http://arxiv.org/abs/2206.02279 (2022).

20. Marshall, S. M. *et al.* Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nat. Commun.* **12**, 3033 (2021).

21. Schmitt-Kopplin, P. *et al.* Systems chemical analytics: introduction to the challenges of chemical complexity analysis. *Faraday Discuss.* **218**, 9–28 (2019).

22. Marshall, S. M., Murray, A. R. G. & Cronin, L. A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **375**, 20160342 (2017).

23. Marshall, S. M., Moore, D. G., Murray, A. R. G., Walker, S. I. & Cronin, L. Formalising the Pathways to Life Using Assembly Spaces. *Entropy* **24**, 884 (2022).

24. Furey, A., Moriarty, M., Bane, V., Kinsella, B. & Lehane, M. Ion suppression; A critical review on causes, evaluation, prevention and applications. *Talanta* **115**, 104–122 (2013).

25. Larkin, P. *Infrared and raman spectroscopy: principles and spectral interpretation*. (Elsevier, 2011).

26. Pupeza, I. *et al.* Field-resolved infrared spectroscopy of biological systems. *Nature* **577**, 52–59 (2020).

27. Orlando, G., Raimondi, D. & F. Vranken, W. Auto-encoding NMR chemical shifts from their native vector space to a residue-level biophysical index. *Nat. Commun.* **10**, 2511 (2019).

28. Baurin, N. *et al.* Design and Characterization of Libraries of Molecular Fragments for Use in NMR Screening against Protein Targets. *J. Chem. Inf. Comput. Sci.* **44**, 2157–2166 (2004).

29. Lau, W. F. *et al.* Design of a multi-purpose fragment screening library using molecular complexity and orthogonal diversity metrics. *J. Comput. Aided Mol. Des.* **25**, 621–636 (2011).

30. Hann, M. M., Leach, A. R. & Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **41**, 856–864 (2001).

31. Bubb, W. A. NMR spectroscopy in the study of carbohydrates: Characterizing the structural complexity. *Concepts Magn. Reson.* **19A**, 1–19 (2003).

32. Marshall, S. M., Moore, D. G., Murray, A. R. G., Walker, S. I. & Cronin, L. Formalising the Pathways to Life Using Assembly Spaces. *Entropy* **24**, 884 (2022).

33. Burger, R. & Bigler, P. DEPTQ: Distorsionless Enhancement by Polarization Transfer Including the Detection of Quaternary Nuclei. *J. Magn. Reson.* **135**, 529–534 (1998).

34. Bigler, P., Kümmerle, R. & Bermel, W. Multiplicity editing including quaternary carbons: improved performance for the13C-DEPTQ pulse sequence. *Magn. Reson. Chem.* **45**, 469–472 (2007).

35. 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: Lab administration, spectra assignment aid and local database. *Magn. Reson. Chem.* **53**, 582–589 (2015).

36. Goodacre, R. *The blind men and the elephant* : challenges in the analysis of complex natural mixtures. *Faraday Discuss.* **218**, 524–539 (2019).

37. Zhao, Y., Kongstad, K. T., Liu, Y., He, C. & Staerk, D. Unraveling the complexity of complex mixtures by combining high-resolution pharmacological, analytical and spectroscopic techniques: antidiabetic constituents in Chinese medicinal plants. *Faraday Discuss.* **218**, 202–218 (2019).

38. Claridge, T. D. W. *High-resolution NMR techniques in organic chemistry*. (Elsevier, 2016).

39. Ballou, E. V., Wood, P. C., Wydeven, T., Lehwalt, M. E. & Mack, R. E. Chemical interpretation of Viking Lander 1 life detection experiment. *Nature* **271**, 644–645 (1978).

40. Mahaffy, P. R. *et al.* The Sample Analysis at Mars Investigation and Instrument Suite. *Space Sci. Rev.* **170**, 401–478 (2012).

41. Waite, J. H. *et al.* Cassini Ion and Neutral Mass Spectrometer: Enceladus Plume Composition and Structure. *Science* **311**, 1419–1422 (2006).

42. https://www.nasa.gov/dragonfly/dragonfly-overview/index.html.

43. Turyshev, S. G. & Toth, V. T. Imaging extended sources with the solar gravitational lens. *Phys. Rev. D* **100**, 084018 (2019).

Supplementary Information for:

Multimodal Techniques for Detecting Alien Life using Assembly Theory and Spectroscopy

Michael Jirasek†[1], Abhishek Sharma†[1], Jessica R. Bame†[1], Nicola Bell[1], Stuart M. Marshall,[1] Cole Mathis[1], Alasdair Macleod,[1] Geoffrey J. T. Cooper[1], Marcel Swart[2,3], Rosa Mollfulleda[2] Leroy Cronin*[1]

[1] School of Chemistry, The University of Glasgow, University Avenue, Glasgow G12 8QQ, UK.

[2] University of Girona, Campus Montilivi (Ciencies), c/M.A. Capmany 69, 17003 Girona Spain

[3]ICREA, Pg. Lluis Companys 23, 08010 Barcelona, Spain

Contents

1 Calculating Assembly Index From Molecular Graph

1.1 Algorithm Description

The assembly index, and associated minimal assembly pathways, are calculated using an algorithm written in the Go programming language. In prior work,[1] the assembly index was calculated using a serial algorithm written in C++, and yielded the "split-branch" assembly index, an approximation that provides a reasonably tight upper bound for the assembly index. The Go algorithm used in this work is a faster algorithm that incorporates concurrency, and can provide the exact assembly index if it can be calculated in a reasonable time. The process can also be terminated early to provide the lowest assembly index found so far, which has been found to be a good approximation for the assembly index in most cases.

The assembly index is calculated by iterating over subgraphs within a molecular graph, and finding duplicates of that subgraph within the remainder of the molecule. For each of the matching subgraphs found an assembly pathway can be represented by a duplicate structure and a remnant structure. The remnant structure comprises the original structure with one duplicate removed, and the other "broken off", which ensures that all structures on an assembly pathway that are duplicated will be first constructed (**Fig. S1**).

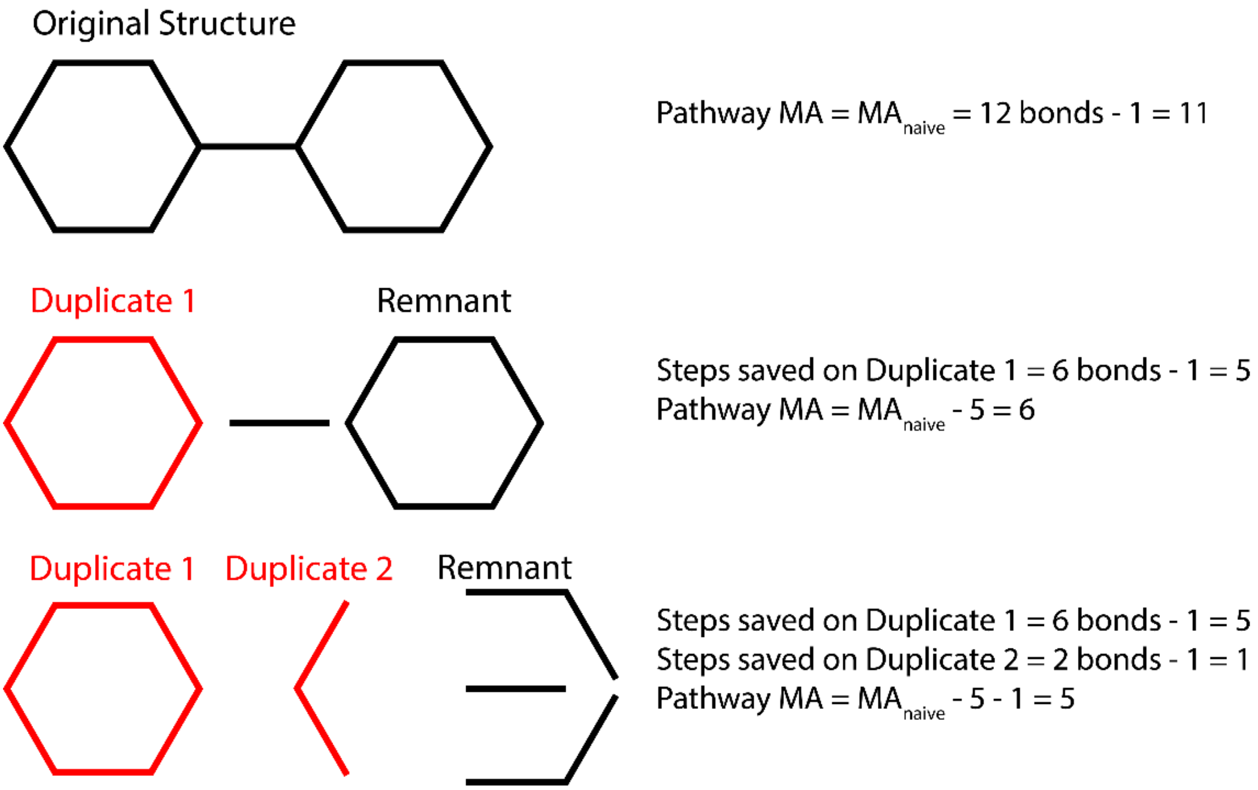

Original Structure
Pathway MA = MA$_{naive}$ = 12 bonds - 1 = 11
Duplicate 1
Remnant
Steps saved on Duplicate 1 = 6 bonds - 1 = 5
Pathway MA = MA$_{naive}$ - 5 = 6
Duplicate 1
Duplicate 2
Remnant
Steps saved on Duplicate 1 = 6 bonds - 1 = 5
Steps saved on Duplicate 2 = 2 bonds - 1 = 1
Pathway MA = MA$_{naive}$ - 5 - 1 = 5

Fig. S1. Illustration of process used by the assembly algorithm to extend an assembly pathway. At each step (top to bottom) a duplicate is found (red) and stored in separately. The remnant is the remaining part of the structure, but with the matching duplicated separated. The process can then be repeated recursively on the remnant.

The process can then be repeated recursively with the remnant structure as an input, which may result in more pathways containing two duplicate structures and a smaller remnant. Thus each pathway is represented by a sequence of duplicated structures and a remnant structure. In order to determine the assembly index, we consider that a molecular graph with N bonds could be constructed in $N-1$ steps by adding one bond at a time (the naive MA, or MA_{naive}). Each duplicate structure of size N_{dup} allows us to add that structure in one step, reducing the number of steps compared to MA_{naive} by $N_{dup}-1$. Thus the MA for a particular pathway is $MA_{naive} - \sum_{dup}(N\_dup - 1)$.

Concurrency is implemented through a worker pool, with each worker iterating over the subgraphs of a particular pathway and placing generated extended pathways into a jobs queue to be picked up and extended by other workers. In order to prevent unbounded resource use, the jobs queue size is limited, and if full a worker will process generated pathways in a depth-first fashion until there is space in the queue, before resuming the breadth first search. The algorithm has some branch and bound methods to reduce the search space (it will not extend pathways that cannot have lower MA than the lowest found so far), and can be terminated early to output the best pathway found so far. The approximation through stopping early has been found to output values at or close to the actual assembly index fairly quickly (**Fig. S2**).

The subgraph iteration process is based on,[2] and the subgraph matching functions are based on processes used in Nauty.[3] The overall algorithm concept is similar to the exact MA algorithm we published previously,[4] but with substantial improvements in terms of performance.

Molecular assembly can be expected to correlate with molecular weight. This is because there are upper and lower bounds for molecular assembly indices that scale with the number of bonds in the molecule. The trivial upper bound for the assembly index relates to pathways where one bond is joined at a time without any leverage of duplication (so the assembly index is equal to the number of bonds minus one). A basic lower bound can be determined by considering that quickest way to increase the size of a structure using an assembly pathway is to take the largest structure created so far and combine it with itself, essentially doubling the size at each step. For example, a structure of 8 bonds cannot be made in less than 3 joining operations, and in general a structure of *N* bonds has a lower bound on the assembly index of $\log_2(N)$. Both these bounds increase with the number of bonds, and since the number of atoms and hence the molecular weight tends to increase with the number of bonds, we can expect the assembly index to increase with the molecular weight.

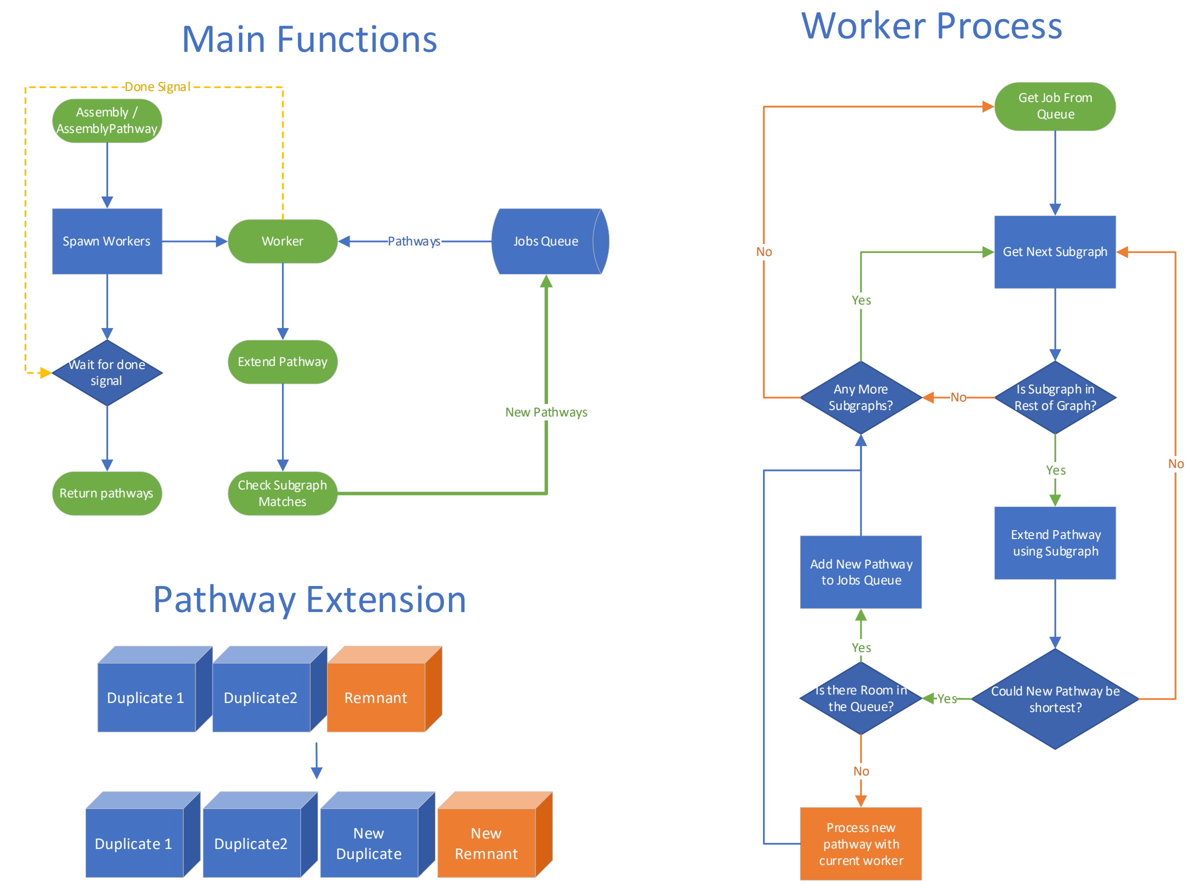


Fig. S2: Outline of algorithm process (top left). Illustration of extension of a single pathway (bottom left). Outline of process for single worker (right).

2 Theoretical calculations

2.1 Database and sampling

The calculation and theoretical basis for Molecular Assembly (MA) calculation are described in detail in our previous work.[1,5] Most of the analysis was performed using Python 3 and Mathematica 12. The Molecular Assembly calculator called *AssemblyGo* was written in GO programming language (https://github.com/croningp/assembly_go). The codes used for processing data and further details can be found at https://github.com/croningp/molecular_complexity.

To study the relationship between the MA and physically measurable properties, we used a previously published database of compounds for which the MA was calculated (~2.2M compounds). In order to address the molecular complexity of organic molecules, and given that we try to address the molecular complexity through carbon-sensitive ^{13}C NMR, hence, a relatively high abundance of carbon is essential. We filtered the compounds to have at least 50% and not more than 85% of the heavy atoms as carbons and must contain at least 4 carbon atoms. Such a filtered database contained *ca.* 0.77 million compounds. The range of the previously calculated Molecular Assembly (called Pathway

Assembly, using the split-branch algorithm) was found in between 3–25. The distribution of MA in the database is not uniform with the highest counts between 8-12 see **Fig. S3**.

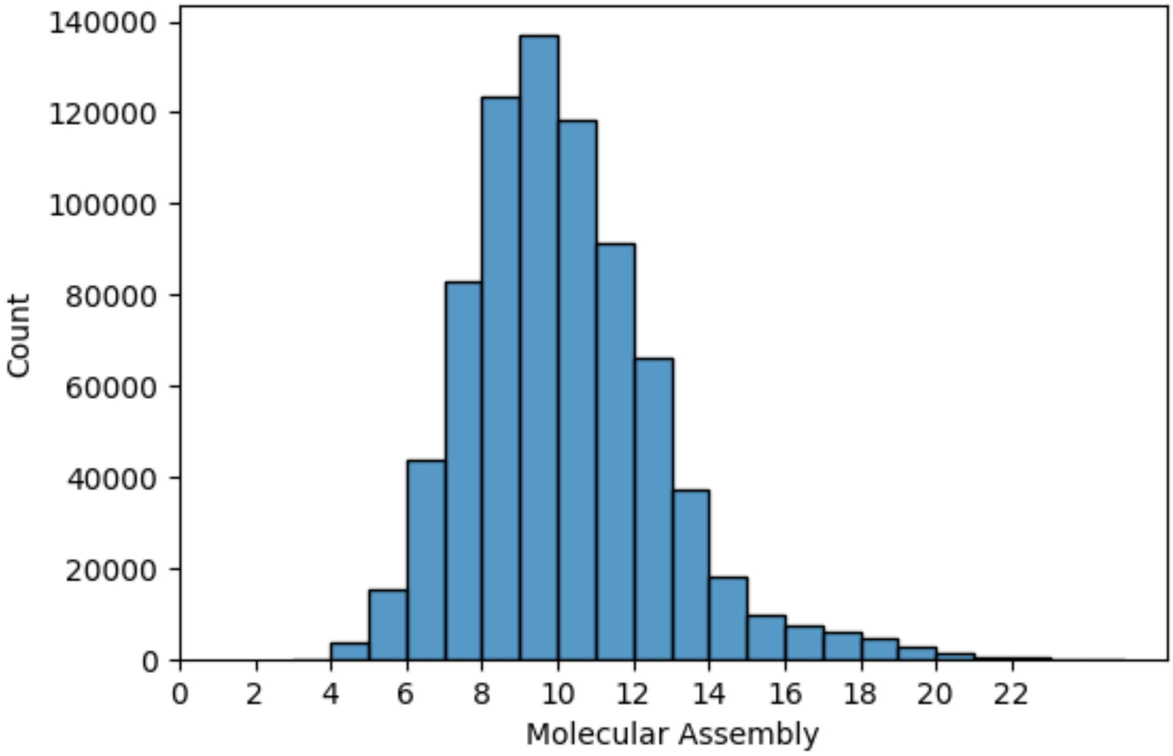


Fig. S3. **Distribution of molecules over the Molecular Assembly from the previous dataset.** The Fig. Shows the histogram of MA (previously called Pathway Assembly) distribution in the available dataset, containing ~0.77 million compounds with 50–85% carbons as heavy atoms, originating from the previously published dataset.[1]

The MA distribution of the compounds filtered from the previously published[1] database reflects synthetic availability (as the compounds originate from the published Reaxys database) and our previous capacity to reliably calculate MA using the split-branch algorithm (compounds for which it was assumed it would be impossible (at the time) to calculate MA reliably were rejected from the database). To assess the characteristic relationship between the MA with the spectroscopic techniques, we sampled 10,000 compounds uniformly across the MA range, up to 629 compounds per MA. The sampling was performed on the subset of which both theoretical NMR and IR data could be calculated using the simulation tools discussed in the later sections. For those compounds, the MA was recalculated using the newly developed assembly algorithm AssemblyGo which provides more accurate estimates at faster timescales, generally leading to estimating the assembly index to be lower by 1 or 2 relative to the original value (previously calculated Pathway Assembly). The distribution of the molecules over the MA range with newly calculated MA is shown in **Fig. S4**.

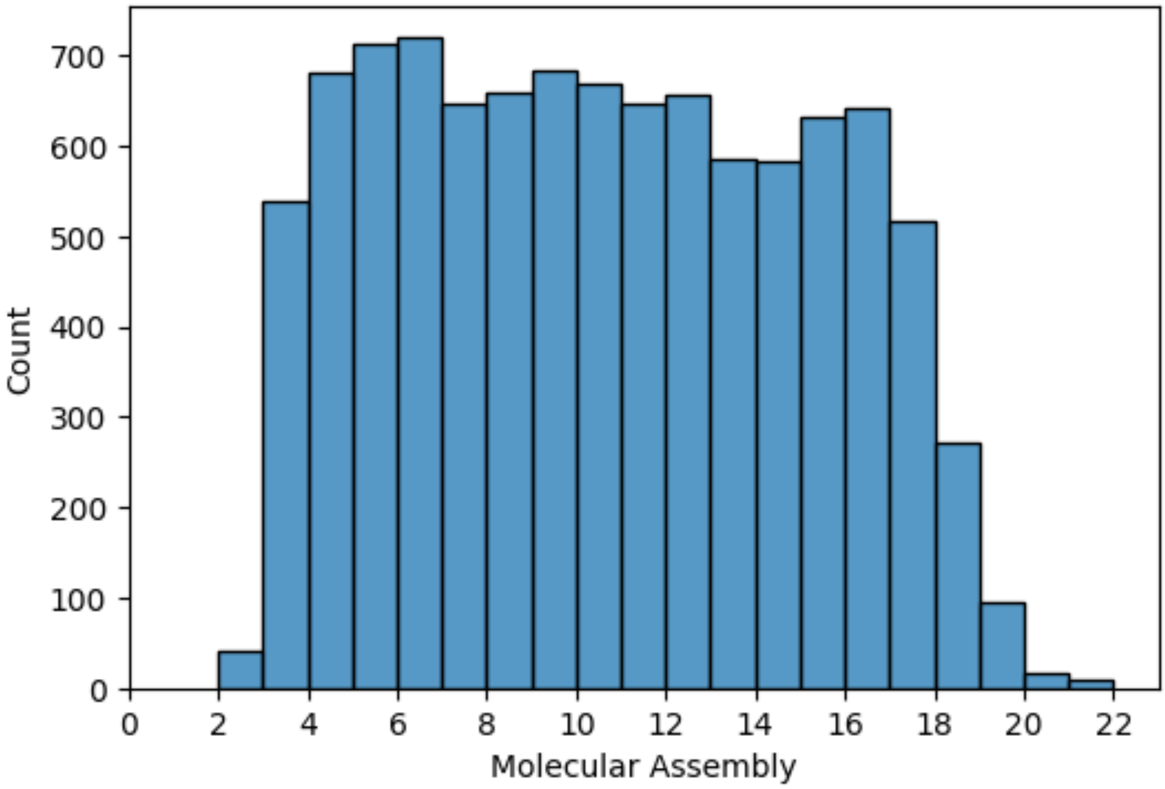


Fig. S4. Histogram of distribution in the sample of 10,000 molecules to cover uniformly the range of MA (recalculated values using a more accurate algorithm).

A representative subset of the molecules in the dataset over the range of MA values is shown in **Fig. S5** on the following page.

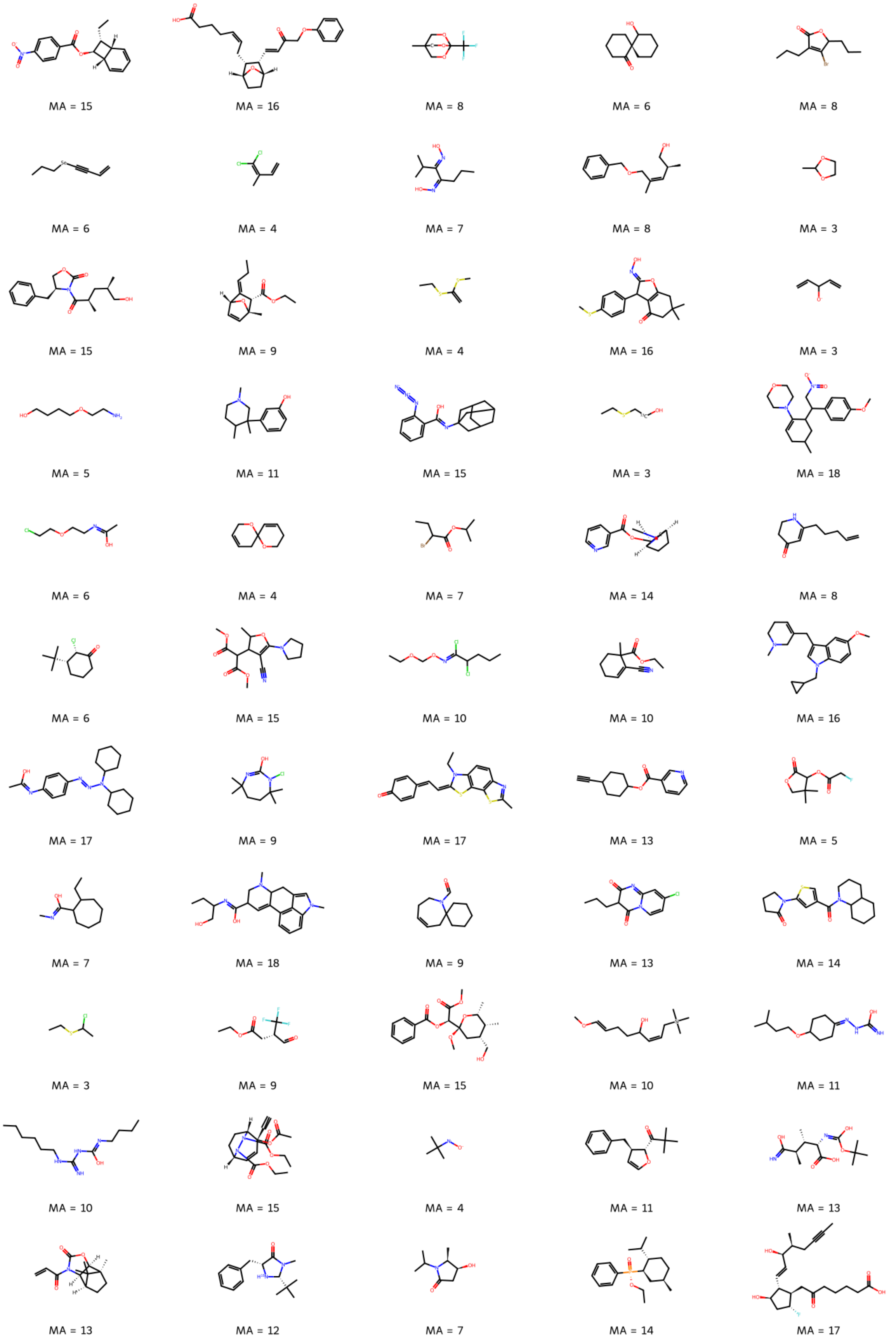


Fig. S5. Example of 55 compounds sampled from the database of 10,000 compounds used in the theoretical study.

2.2 NMR Prediction

The ^{13}C NMR spectra were predicted using *nmrshiftdb2* tool.[6] The corresponding chemical shifts were grouped by the type of carbon (primary (CH_3), secondary (CH_2), tertiary (CH) and quarternary (C)). Such sorting was performed by Python script using the rdkit[7] tool to estimate the number of hydrogen atoms attached. Further, the number of chemical shifts was binned, applying the minimum 0.5 ppm chemical shifts difference (i.e. resonances within the 0.5 ppm were considered as a single peak for analysis).

The importance of an actual ^{13}C NMR measure/prediction instead of the sole counting of the number of chemically non-equivalent carbons in structure can be demonstrated on a large dataset of ~1.1 million compounds (allowing all compounds with more than 4 carbons and no constraints on the C content). This set was analysed by both NMR prediction, as well as by counting the number of nonequivalent carbons (for simplicity, the carbons were not classified by the type) (**Fig. S6**). Also, note that considered assembly index values are based on the old database that used the previous algorithm which is relatively less accurate. The potential outliers deviating from the linear trends highlight the utility of the actual NMR measure (as an oriented oligomer possesses plenty of non-equivalent carbons, yet of very similar chemical shift).

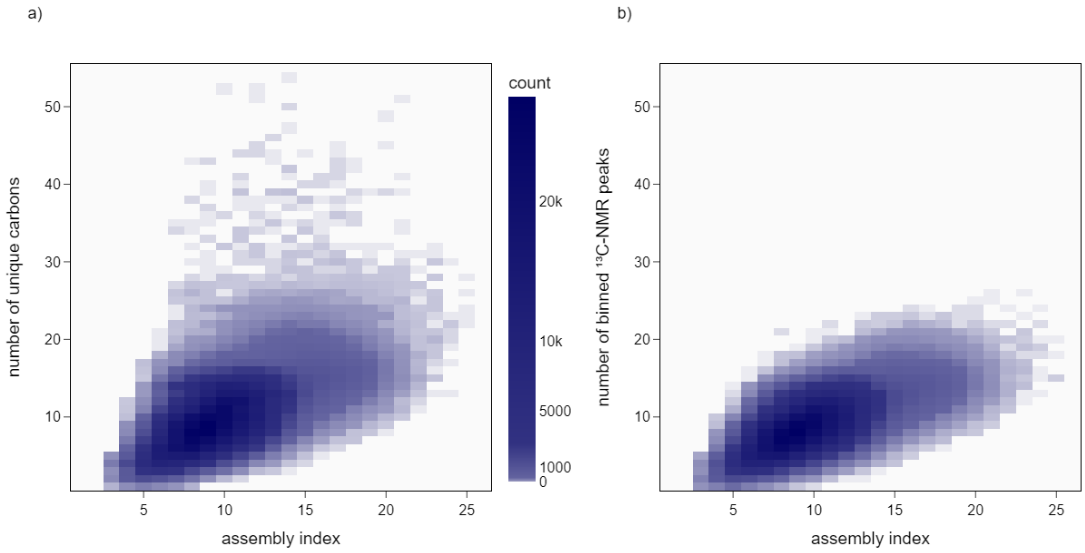
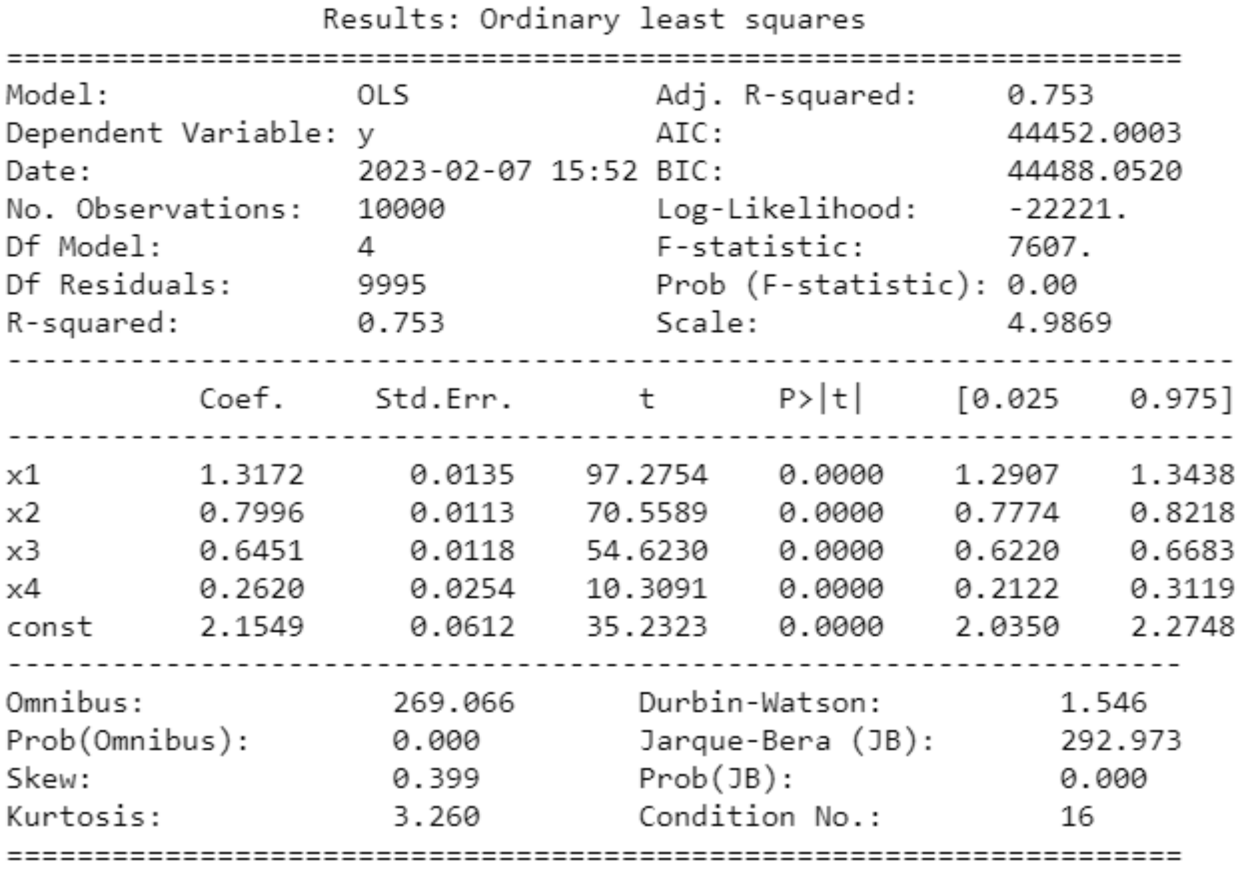


Fig. S6. **Analysis of *ca.* 1.1 million compounds.** a) Molecular Assembly (assembly index) *vs.* the number of unique carbons. b) MA *vs.* the number of predicted 0.5 ppm binned ^{13}C NMR resonances. Note that the z-axis (histogram count) is scaled logarithmically with base 10 to emphasise even the very uncommon cases.

On the sample of 10,000 molecules, whose MA was recalculated using the new algorithm, we used multivariate fit of the number of 0.5 ppm binned resonances of C, CH, CH_2 and CH_3 carbon resonances. For the fit, the *statsmodels.api.OLS* module in Python was used (**Fig. S7**).[8]



```
                     Results: Ordinary least squares
===================================================================
Model:               OLS              Adj. R-squared:     0.753
Dependent Variable:  y                AIC:                44452.0003
Date:                2023-02-07 15:52 BIC:                44488.0520
No. Observations:    10000            Log-Likelihood:     -22221.
Df Model:            4                F-statistic:        7607.
Df Residuals:        9995             Prob (F-statistic): 0.00
R-squared:           0.753            Scale:              4.9869
----------------------------------------------------------------------
           Coef.     Std.Err.      t        P>|t|     [0.025    0.975]
----------------------------------------------------------------------
x1         1.3172      0.0135    97.2754    0.0000    1.2907    1.3438
x2         0.7996      0.0113    70.5589    0.0000    0.7774    0.8218
x3         0.6451      0.0118    54.6230    0.0000    0.6220    0.6683
x4         0.2620      0.0254    10.3091    0.0000    0.2122    0.3119
const      2.1549      0.0612    35.2323    0.0000    2.0350    2.2748
-------------------------------------------------------------------
Omnibus:               269.066       Durbin-Watson:          1.546
Prob(Omnibus):         0.000         Jarque-Bera (JB):       292.973
Skew:                  0.399         Prob(JB):               0.000
Kurtosis:              3.260         Condition No.:          16
===================================================================
```

Fig. S7. Print output from the multivariate fit of MA = x_1×C + x_2×CH + x_3×CH_2 + x_4×CH_3 + *const.*; using *statsmodels.api.OLS* in python.[8]

The best prediction of MA based on the NMR data is given by:

$$\mathrm{MA} = 1.32 \times \mathrm{C} + 0.80 \times \mathrm{CH} + 0.65 \times \mathrm{CH_2} + 0.26 \times \mathrm{CH_3} + 2.15 \qquad \textbf{(1)}$$

where C, CH, CH_2 and CH_3 are the number of calculated unique (binned with 0.5 ppm resolution) ^{13}C resonances corresponding to carbons with 0, 1, 2 and 3 attached hydrogens, respectively. The distribution of MA vs. NMR-predicted MA is visualised as a histogram is shown in **Fig. S8**.

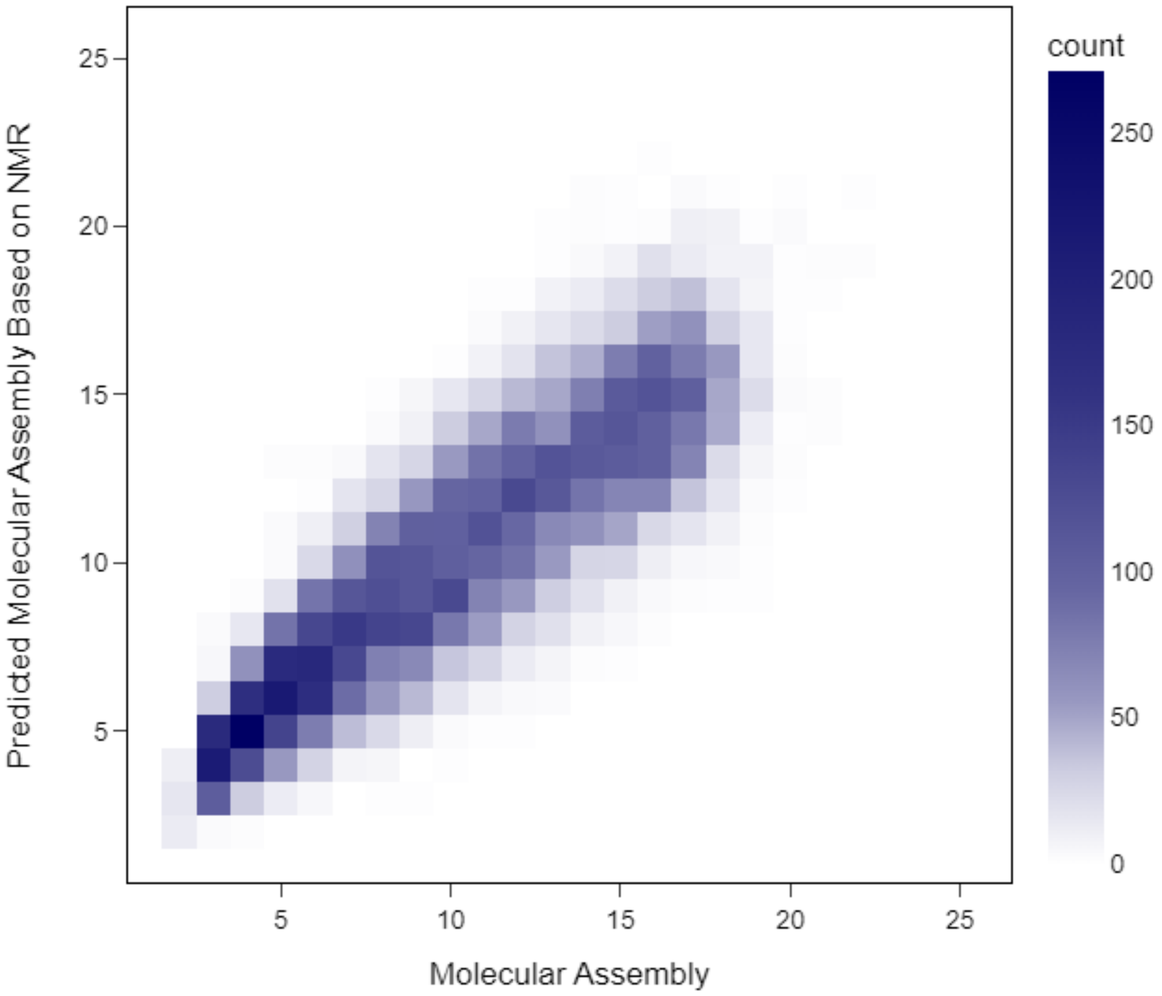
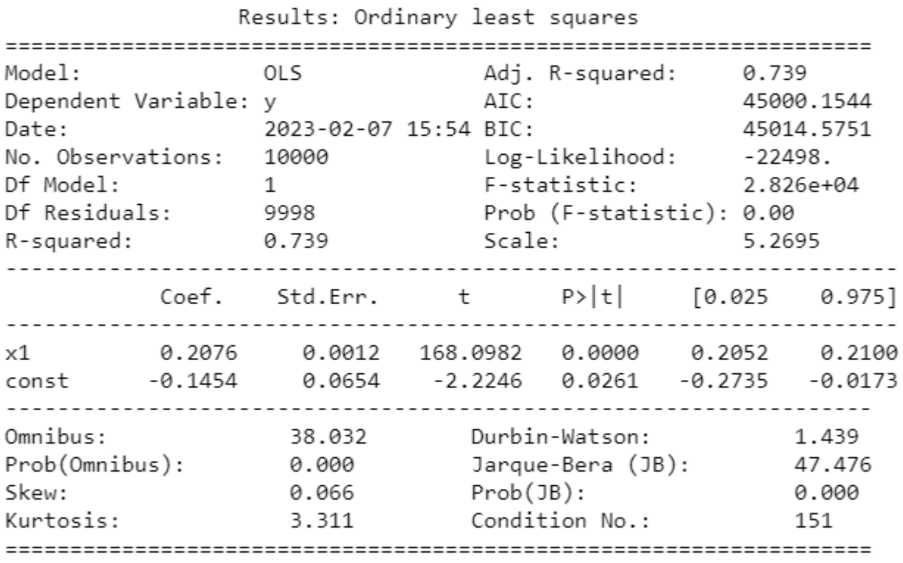


Fig. S8. Histogram of predicted MA (based on the **Eq. 1**) *vs.* MA on 10,000 compounds sample.

2.3 Infrared Spectroscopy – xTB simulations

To predict the IR spectra of the sampled molecules, we have used the xTB-service tool[9,10] for faster prediction over a large dataset. The provided Python interface[11] was used with the default setting, using 100 seconds as timeout for the geometry optimisation using the GFNFF forcefield. The default gaussian broadening was not applied to the observed intensity. Using the default threshold checks, calculated spectra assumed for the interpretation were based on the molecule with no large imaginary frequency (set as maximum $i{\cdot}10\ \mathrm{cm}^{-1}$, although many structures possess small imaginary frequencies). The peaks in the range of 400–1500 cm^{-1} were counted with a threshold of 0.0005 $(\mathrm{D/Å})^2 \cdot \mathrm{amu}^{-1}$ to not consider signals with 0 oscillatory strength and binned together peaks within 2 cm^{-1}. The coefficients for the simple linear function of the number of IR peaks were fit using the *statsmodels.api.OLS* module in python (**Fig. S9**).[8]



```
                   Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.739
Dependent Variable: y                AIC:                45000.1544
Date:               2023-02-07 15:54 BIC:                45014.5751
No. Observations:   10000            Log-Likelihood:     -22498.
Df Model:           1                F-statistic:        2.826e+04
Df Residuals:       9998             Prob (F-statistic): 0.00
R-squared:          0.739            Scale:              5.2695
--------------------------------------------------------------------
            Coef.    Std.Err.      t       P>|t|    [0.025    0.975]
--------------------------------------------------------------------
x1          0.2076     0.0012   168.0982   0.0000    0.2052    0.2100
const      -0.1454     0.0654    -2.2246   0.0261   -0.2735   -0.0173
------------------------------------------------------------------
Omnibus:              38.032        Durbin-Watson:           1.439
Prob(Omnibus):        0.000         Jarque-Bera (JB):        47.476
Skew:                 0.066         Prob(JB):                0.000
Kurtosis:             3.311         Condition No.:           151
==================================================================
```

Fig. S9. Print output from the fit of MA = $x_1 \times n_{\text{peaks}}$ + *const.*; using *statsmodels.api.OLS* in python.[8]

The best prediction of MA based on the xTB-based IR predicted data is thus:

$$\mathrm{MA} = 0.21 \times n_{\mathrm{IR\_peaks}} - 0.15 \qquad \textbf{(2)}$$

where $n_{\text{IR\_peaks}}$ is the number of IR peaks in the region of 400–1500 cm^{-1} with intensity above 0.0005 (D/Å)^2 · amu^{-1}. The distribution of MA vs. IR-predicted MA is visualised as a histogram in **Fig. S10**.

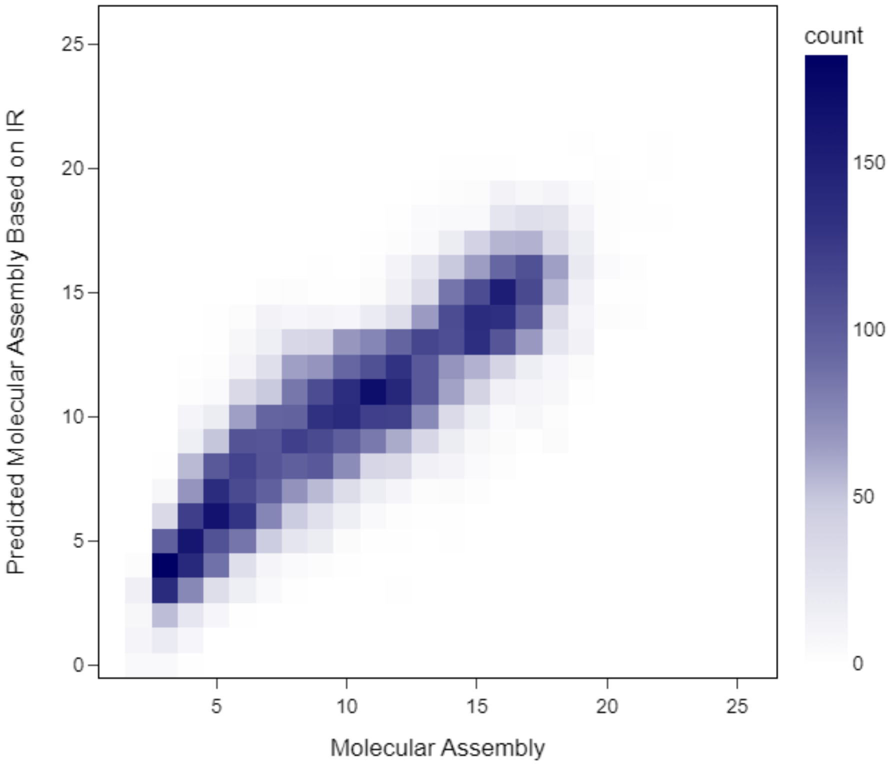
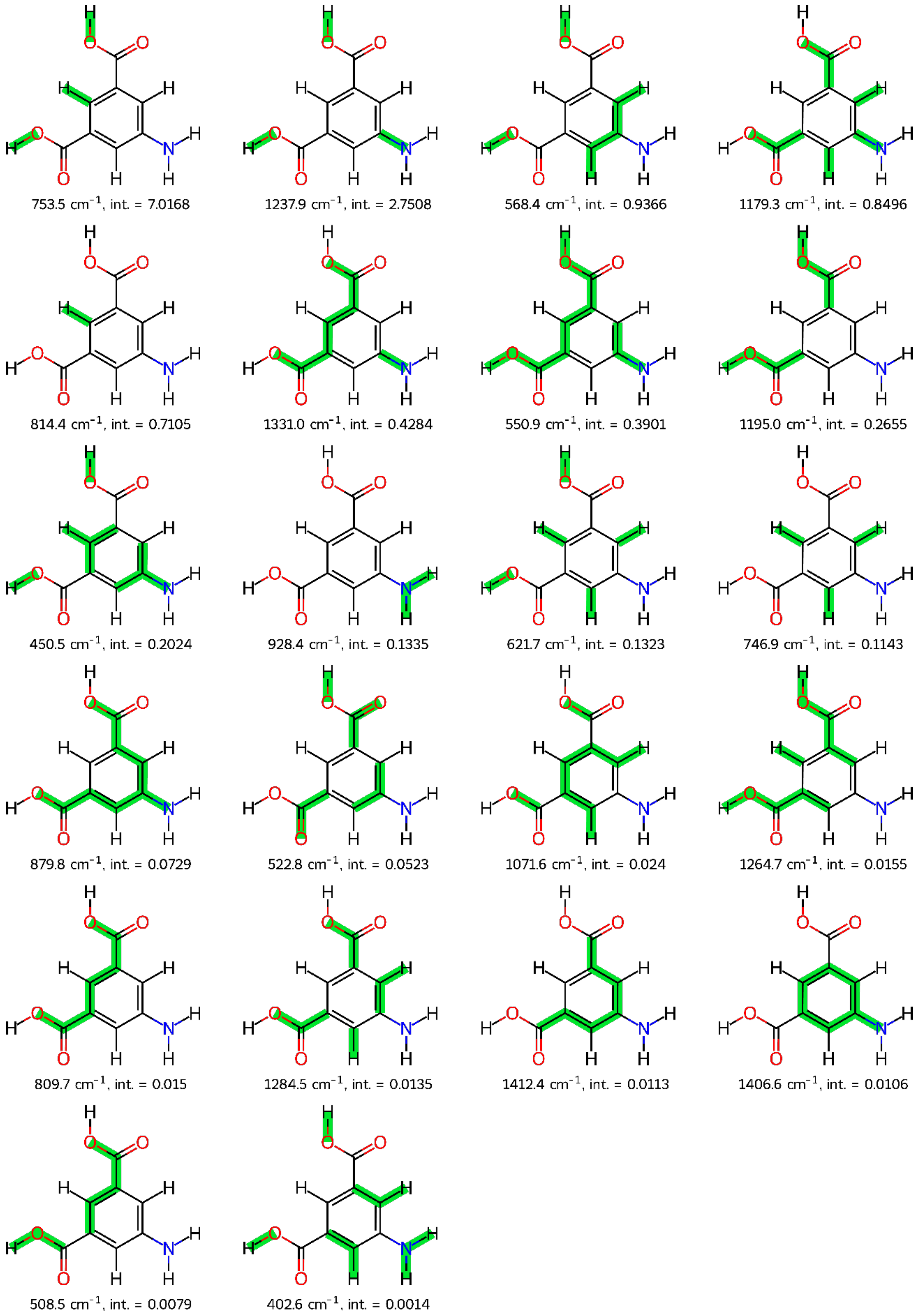


Fig. S10. Histogram of predicted MA (based on the **Eq. 2**) *vs.* MA on 10,000 compounds dataset.

Our general hypothesis is that modes in the IR fingerprint region could be associated largely with collective motions, involving bonds from the various subgraphs of the whole structure. Therefore, from number of the total modes in the fingerprint region the overall molecular complexity could be inferred. To illustrate that on a simple and a complex molecule, vibrational modes in the fingerprint region (400–1500 cm^{-1}) above the set intensity threshold of 0.0005 (D/Å)^2 · amu^{-1} for chemical structures of 5-aminoisopthalic acid (**Fig. S11**.) and quinine (

Fig. S12–**Fig. S15**) are visualised. On the molecular structure, bonds involved it the vibrational modes are highlighted.



753.5 cm^{-1}, int. = 7.0168

1237.9 cm^{-1}, int. = 2.7508

568.4 cm^{-1}, int. = 0.9366

1179.3 cm^{-1}, int. = 0.8496

814.4 cm^{-1}, int. = 0.7105

1331.0 cm^{-1}, int. = 0.4284

550.9 cm^{-1}, int. = 0.3901

1195.0 cm^{-1}, int. = 0.2655

450.5 cm^{-1}, int. = 0.2024

928.4 cm^{-1}, int. = 0.1335

621.7 cm^{-1}, int. = 0.1323

746.9 cm^{-1}, int. = 0.1143

879.8 cm^{-1}, int. = 0.0729

522.8 cm^{-1}, int. = 0.0523

1071.6 cm^{-1}, int. = 0.024

1264.7 cm^{-1}, int. = 0.0155

809.7 cm^{-1}, int. = 0.015

1284.5 cm^{-1}, int. = 0.0135

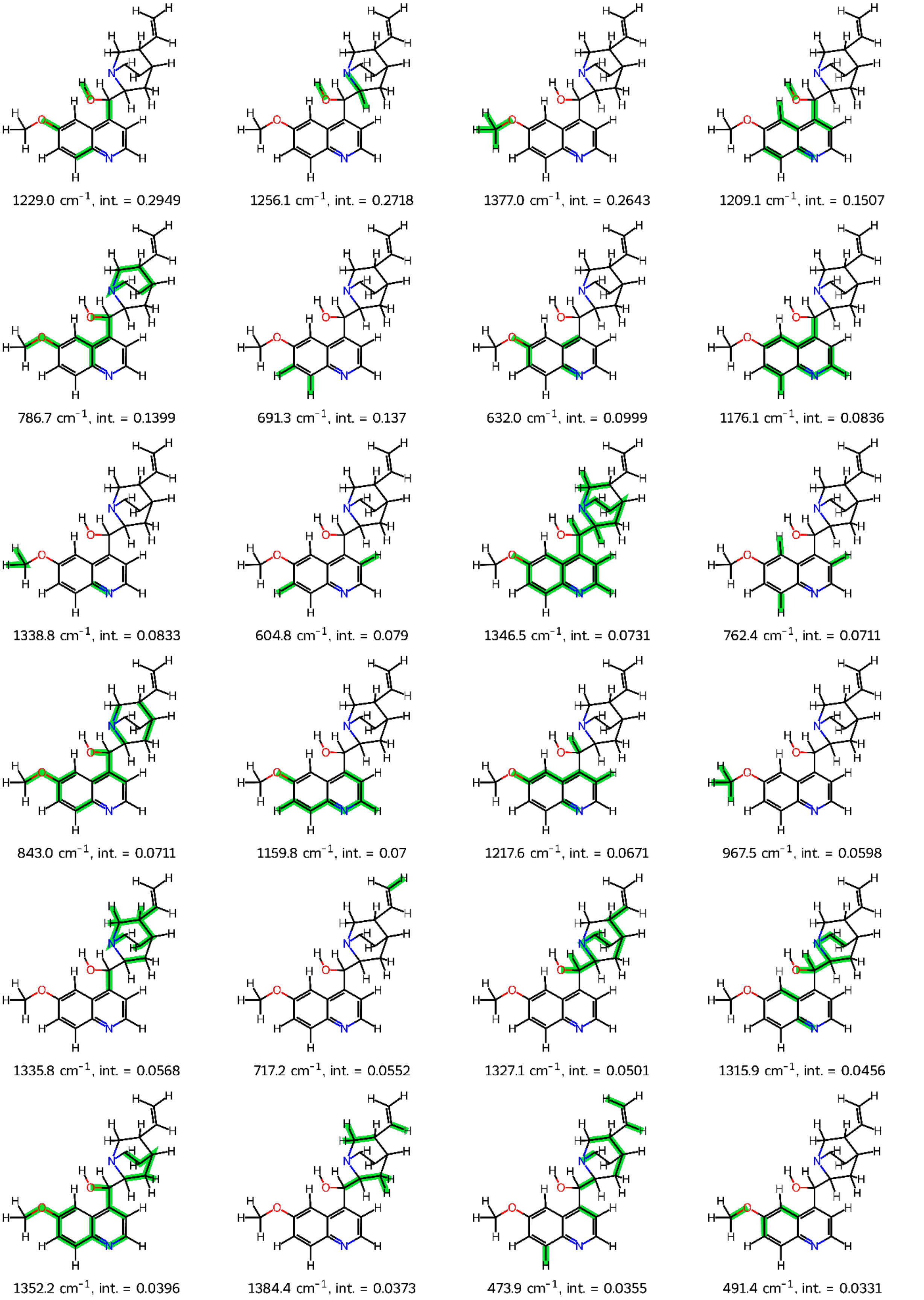
1412.4 cm^{-1}, int. = 0.0113

1406.6 cm^{-1}, int. = 0.0106

508.5 cm^{-1}, int. = 0.0079

402.6 cm^{-1}, int. = 0.0014

Fig. S11. Example of all vibrational bands of 5-aminoisopthalic acid in the fingerprint region demonstrating its collective-motion nature. Vibrational modes are ordered by intensity as calculated by xTB.



1229.0 cm^{-1}, int. = 0.2949

1256.1 cm^{-1}, int. = 0.2718

1377.0 cm^{-1}, int. = 0.2643

1209.1 cm^{-1}, int. = 0.1507

786.7 cm^{-1}, int. = 0.1399

691.3 cm^{-1}, int. = 0.137

632.0 cm^{-1}, int. = 0.0999

1176.1 cm^{-1}, int. = 0.0836

1338.8 cm^{-1}, int. = 0.0833

604.8 cm^{-1}, int. = 0.079

1346.5 cm^{-1}, int. = 0.0731

762.4 cm^{-1}, int. = 0.0711

843.0 cm^{-1}, int. = 0.0711

1159.8 cm^{-1}, int. = 0.07

1217.6 cm^{-1}, int. = 0.0671

967.5 cm^{-1}, int. = 0.0598

1335.8 cm^{-1}, int. = 0.0568

717.2 cm^{-1}, int. = 0.0552

1327.1 cm^{-1}, int. = 0.0501

1315.9 cm^{-1}, int. = 0.0456

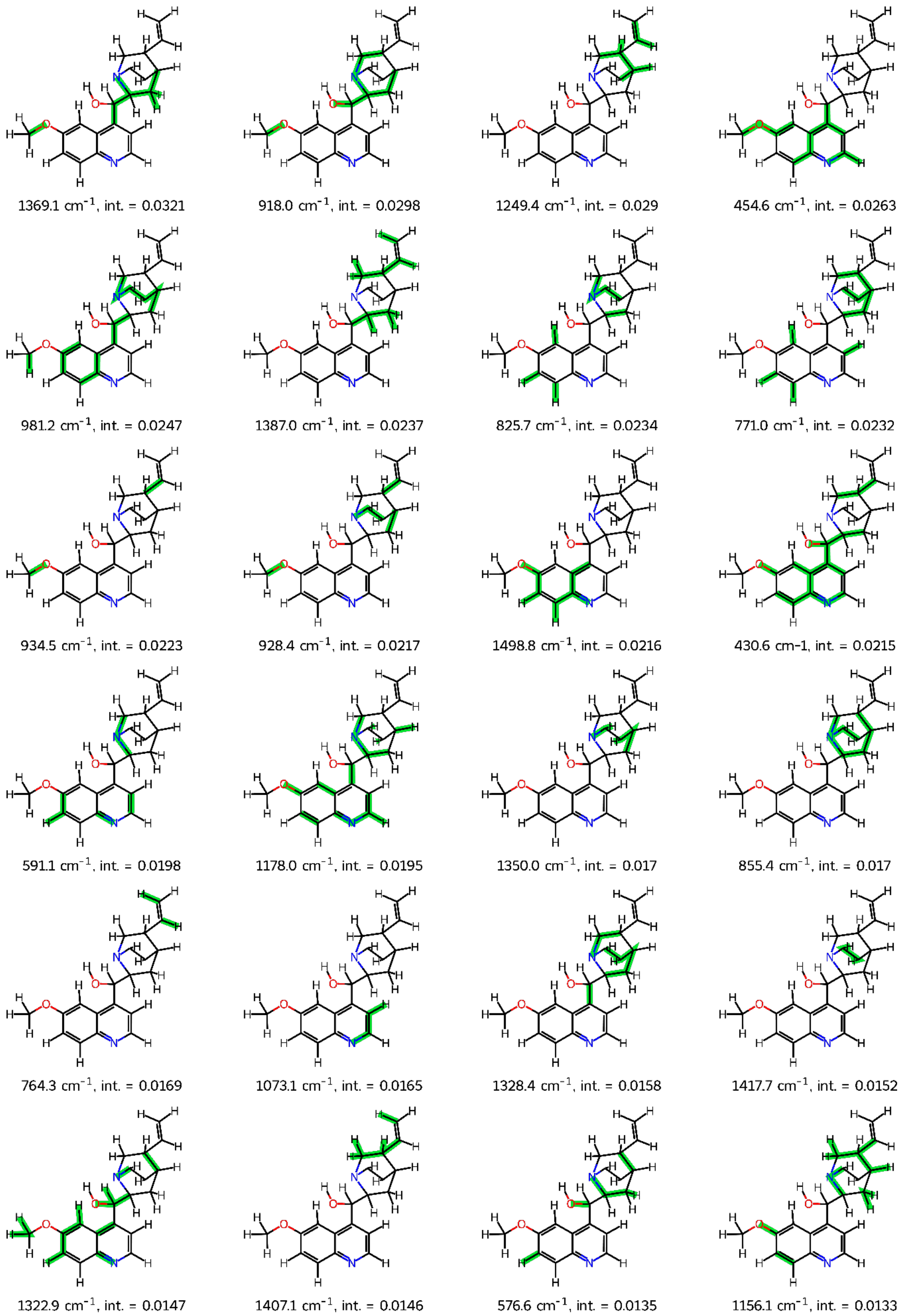
1352.2 cm^{-1}, int. = 0.0396

1384.4 cm^{-1}, int. = 0.0373

473.9 cm^{-1}, int. = 0.0355

491.4 cm^{-1}, int. = 0.0331

Fig. S12. Example of all vibrational bands of quinine in the fingerprint region demonstrating its collective-motion nature. Vibrational modes are ordered by intensity as calculated by xTB. (part 1)



1369.1 cm^{-1}, int. = 0.0321

918.0 cm^{-1}, int. = 0.0298

1249.4 cm^{-1}, int. = 0.029

454.6 cm^{-1}, int. = 0.0263

981.2 cm^{-1}, int. = 0.0247

1387.0 cm^{-1}, int. = 0.0237

825.7 cm^{-1}, int. = 0.0234

771.0 cm^{-1}, int. = 0.0232

934.5 cm^{-1}, int. = 0.0223

928.4 cm^{-1}, int. = 0.0217

1498.8 cm^{-1}, int. = 0.0216

430.6 cm-1, int. = 0.0215

591.1 cm^{-1}, int. = 0.0198

1178.0 cm^{-1}, int. = 0.0195

1350.0 cm^{-1}, int. = 0.017

855.4 cm^{-1}, int. = 0.017

764.3 cm^{-1}, int. = 0.0169

1073.1 cm^{-1}, int. = 0.0165

1328.4 cm^{-1}, int. = 0.0158

1417.7 cm^{-1}, int. = 0.0152

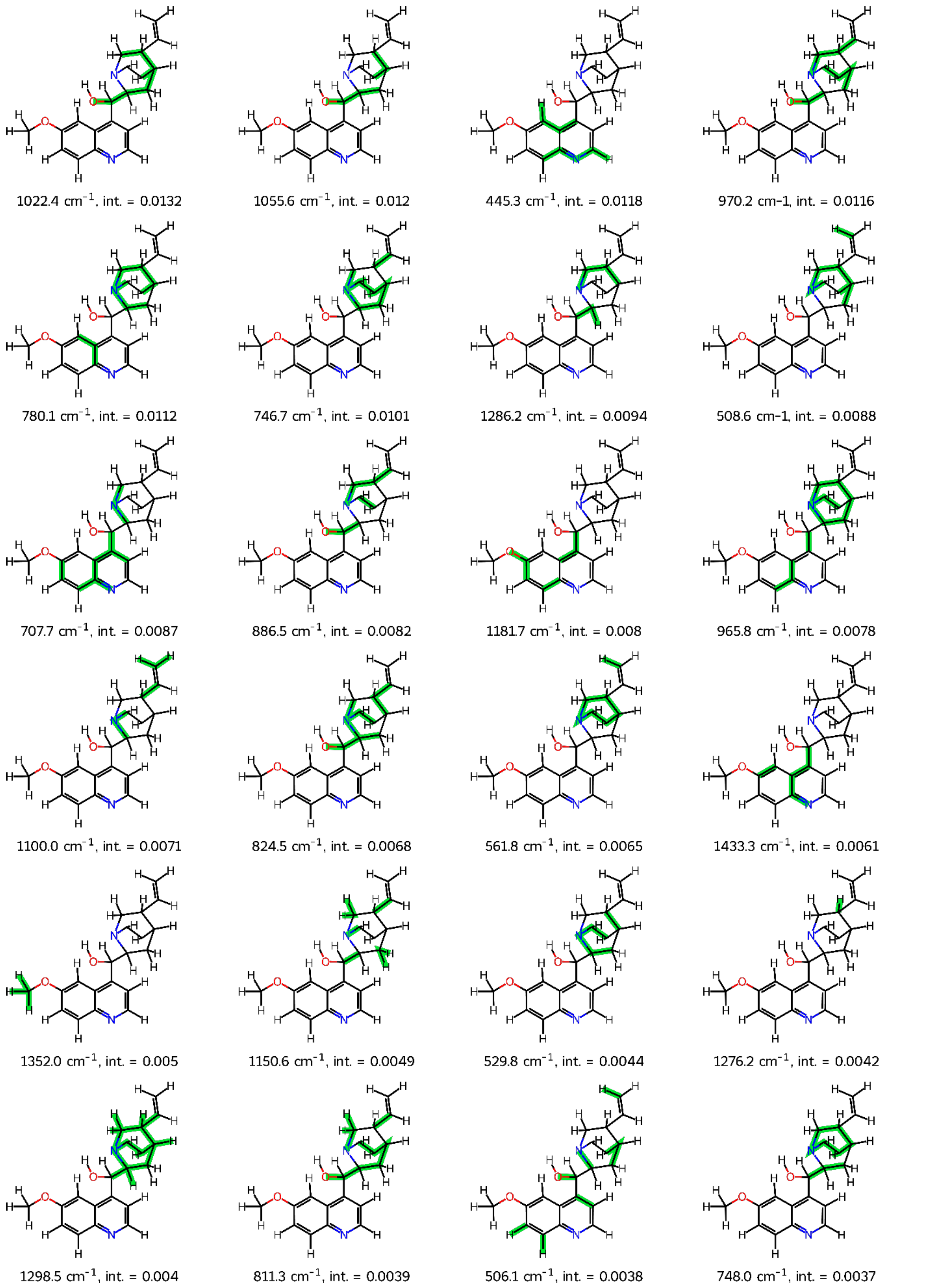
1322.9 cm^{-1}, int. = 0.0147

1407.1 cm^{-1}, int. = 0.0146

576.6 cm^{-1}, int. = 0.0135

1156.1 cm^{-1}, int. = 0.0133

Fig. S13. Example of all vibrational bands of quinine in the fingerprint region demonstrating its collective-motion nature. Vibrational modes are ordered by intensity as calculated by xTB. (part 2)



1022.4 cm^{-1}, int. = 0.0132

1055.6 cm^{-1}, int. = 0.012

445.3 cm^{-1}, int. = 0.0118

970.2 cm-1, int. = 0.0116

780.1 cm^{-1}, int. = 0.0112

746.7 cm^{-1}, int. = 0.0101

1286.2 cm^{-1}, int. = 0.0094

508.6 cm-1, int. = 0.0088

707.7 cm^{-1}, int. = 0.0087

886.5 cm^{-1}, int. = 0.0082

1181.7 cm^{-1}, int. = 0.008

965.8 cm^{-1}, int. = 0.0078

1100.0 cm^{-1}, int. = 0.0071

824.5 cm^{-1}, int. = 0.0068

561.8 cm^{-1}, int. = 0.0065

1433.3 cm^{-1}, int. = 0.0061

1352.0 cm^{-1}, int. = 0.005

1150.6 cm^{-1}, int. = 0.0049

529.8 cm^{-1}, int. = 0.0044

1276.2 cm^{-1}, int. = 0.0042

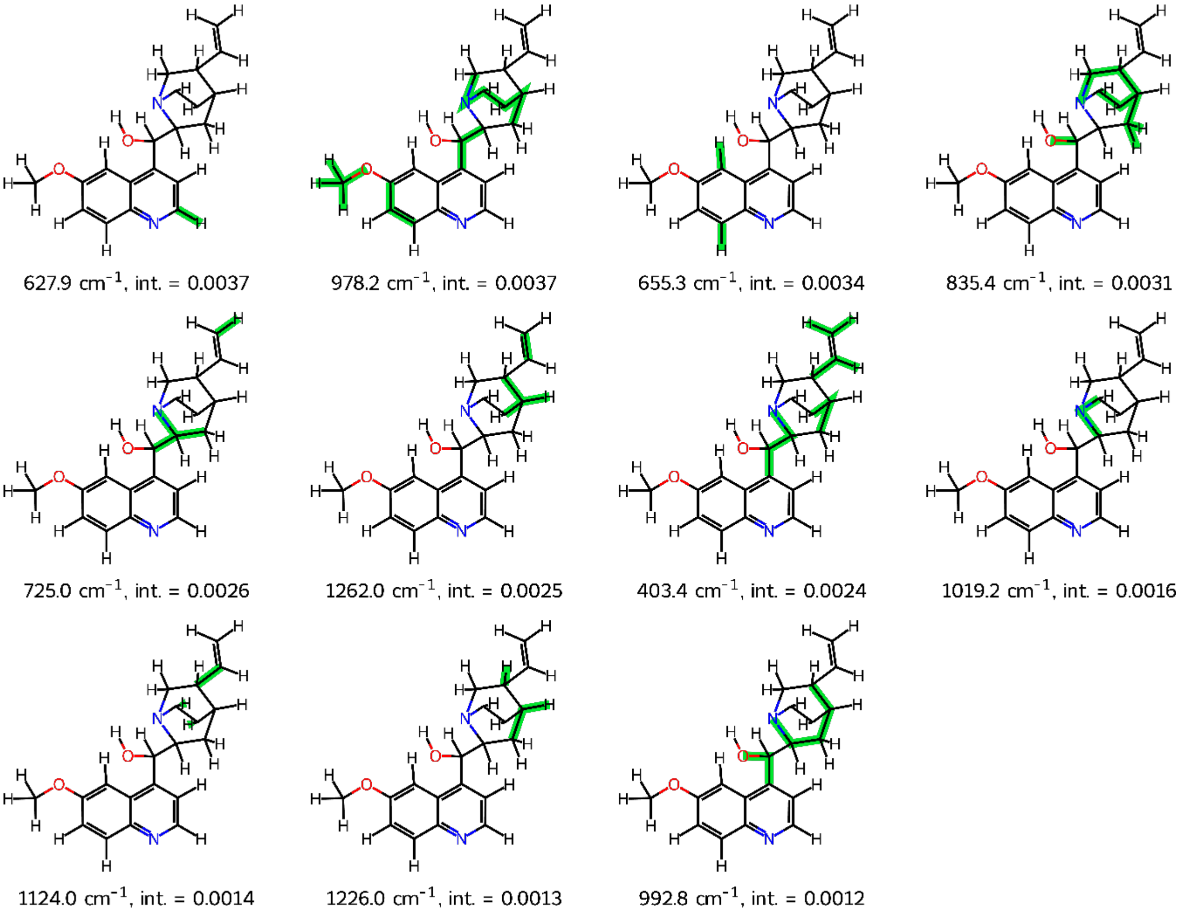
1298.5 cm^{-1}, int. = 0.004

811.3 cm^{-1}, int. = 0.0039

506.1 cm^{-1}, int. = 0.0038

748.0 cm^{-1}, int. = 0.0037

Fig. S14. Example of all vibrational bands of quinine in the fingerprint region demonstrating its collective-motion nature. Vibrational modes are ordered by intensity as calculated by xTB. (part 3)



627.9 cm^{-1}, int. = 0.0037

978.2 cm^{-1}, int. = 0.0037

655.3 cm^{-1}, int. = 0.0034

835.4 cm^{-1}, int. = 0.0031

725.0 cm^{-1}, int. = 0.0026

1262.0 cm^{-1}, int. = 0.0025

403.4 cm^{-1}, int. = 0.0024

1019.2 cm^{-1}, int. = 0.0016

1124.0 cm^{-1}, int. = 0.0014

1226.0 cm^{-1}, int. = 0.0013

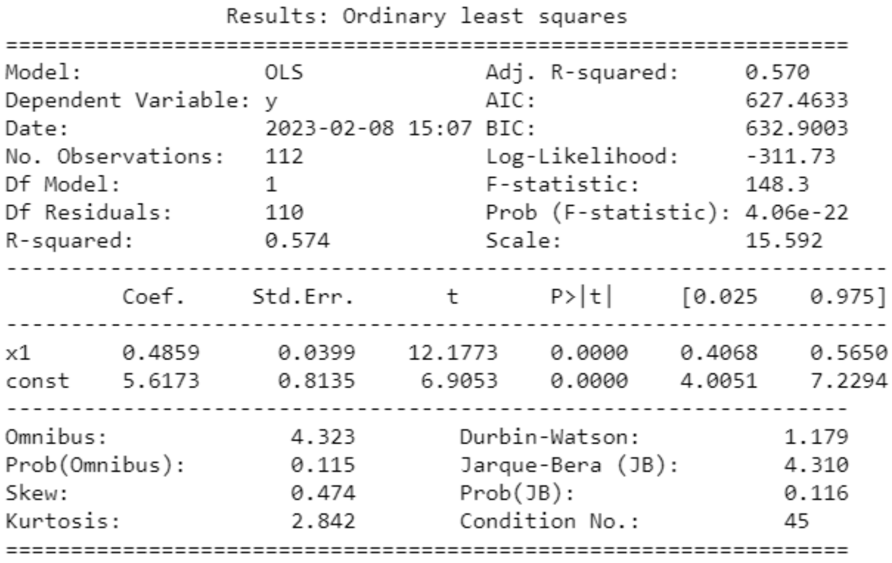
992.8 cm^{-1}, int. = 0.0012

Fig. S15. Example of all vibrational bands of quinine in the fingerprint region demonstrating its collective-motion nature. Vibrational modes are ordered by intensity as calculated by xTB. (part 4)

2.4 Infrared Spectroscopy – DFT simulations

To validate potential inaccuracies in the predicted frequency spectra using the semiempirical method, a detailed rigorous analysis over a limited set of 101 compounds was performed. The theoretical approach has been used as described in the work by Swart and colleagues.[12] Detailed quantum chemical simulations were performed using Amsterdam Density Functional (ADF 2017)[13] software. In this study, QUILD[14] (Quantum-regions interconnected by Local descriptions) program was used with delocalized coordinates for optimization of equilibrium structures until the maximum gradient component was less than 10^{-4} a.u. Energies, gradients and Hessians for vibrational frequencies including Raman intensities were calculated using BP86-D3[15–17] with a triple/double-zeta valence plus polarization basis set (TZP for metals, DZP for other elements). In all cases, these calculations included solvation effects through the COSMO[18] dielectric continuum model with appropriate parameters for solvent, and scalar relativistic corrections through Zeroth Order Regular Approximation (ZORA).[11]

The number of peaks in the fingerprint region (400–1500 cm^{-1}) above an intensity threshold (25 km·mol^{-1}) was found to be 0.76 (**Fig. S16**). The histogram of calculated MA *vs*. the expected is depicted in **Fig. S17**.



```
                     Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.570
Dependent Variable: y                AIC:                627.4633
Date:               2023-02-08 15:07 BIC:                632.9003
No. Observations:   112              Log-Likelihood:     -311.73
Df Model:           1                F-statistic:        148.3
Df Residuals:       110              Prob (F-statistic): 4.06e-22
R-squared:          0.574            Scale:              15.592
--------------------------------------------------------------------
         Coef.     Std.Err.       t        P>|t|      [0.025    0.975]
--------------------------------------------------------------------
x1       0.4859      0.0399    12.1773    0.0000    0.4068    0.5650
const    5.6173      0.8135     6.9053    0.0000    4.0051    7.2294
-----------------------------------------------------------------
Omnibus:               4.323        Durbin-Watson:          1.179
Prob(Omnibus):         0.115        Jarque-Bera (JB):       4.310
Skew:                  0.474        Prob(JB):               0.116
Kurtosis:              2.842        Condition No.:          45
=================================================================
```

Fig. S16. Print output from the fit of MA = $x_1 \times n_{\mathrm{IR\_peaks}}$ + *const.*; using *statsmodels.api.OLS* in Python.[8]

List of chemical structures of all compounds used for the DFT study is in **Fig. S18–Fig. S20**.

The best model for infering the MA based on the DFT-predicted IR spectra is thus:

$$\mathrm{MA} = 0.49 \times n_{\mathrm{IR\_peaks}} + 5.6 \tag{3}$$

where $n_{\mathrm{IR\_peaks}}$ is the number of IR peaks in the region of 400–1500 cm^{-1} with intensity above 25 km·mol^{-1}.

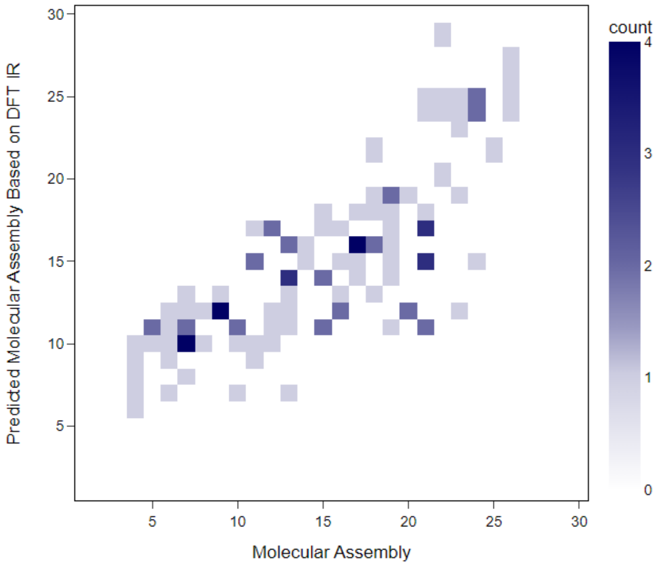


Fig. S17. Histogram of predicted MA (based on the **Eq. 3**) *vs.* MA on 112 compounds dataset of DFT calculated IR peaks in the range of 400–1500 cm^{-1} above intensity 25 km·mol^{-1}.

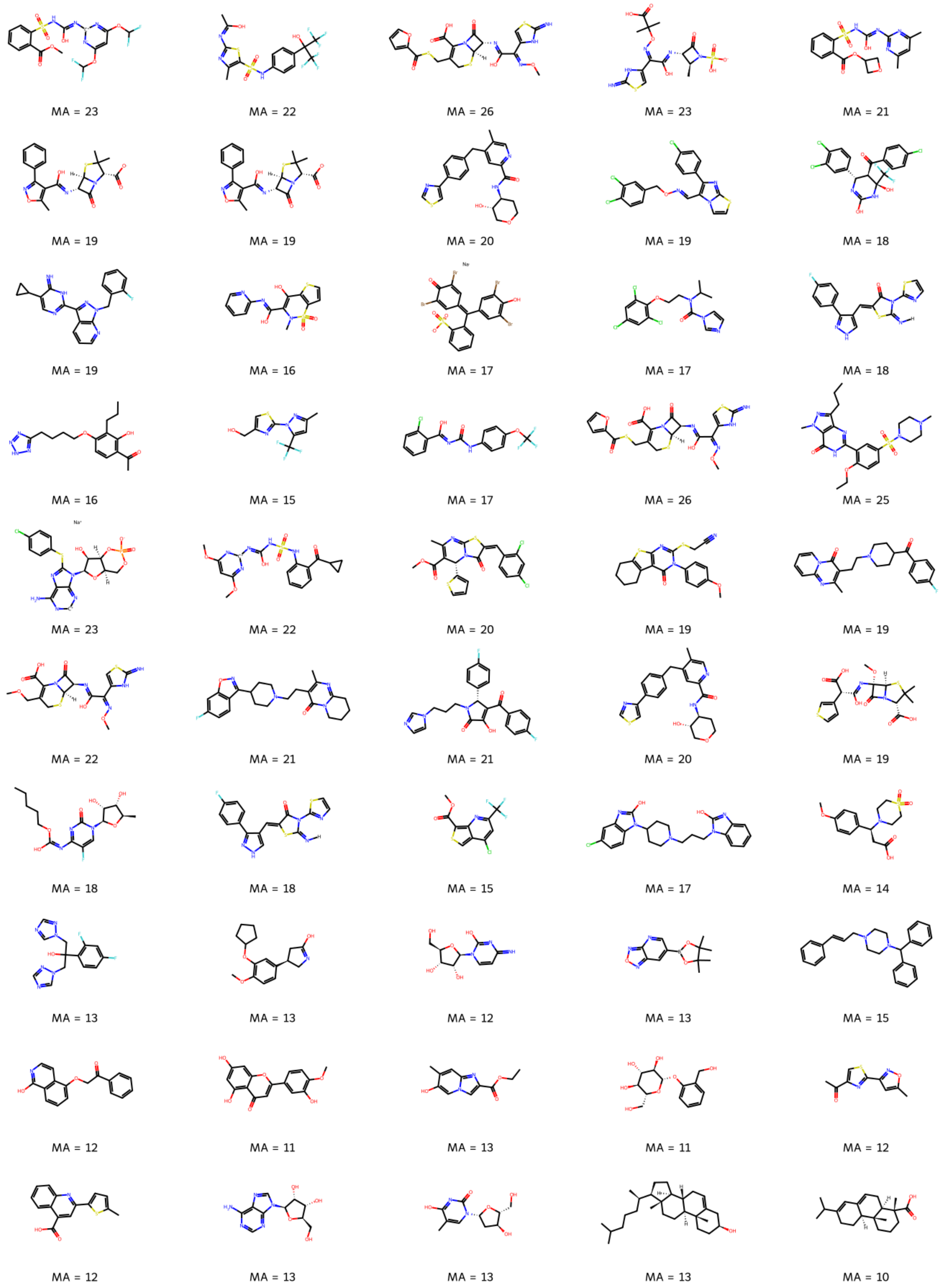


Fig. S18. Molecular structures with calculated molecular assembly (MA) were used in the DFT-calculated IR study (**Part 1**).

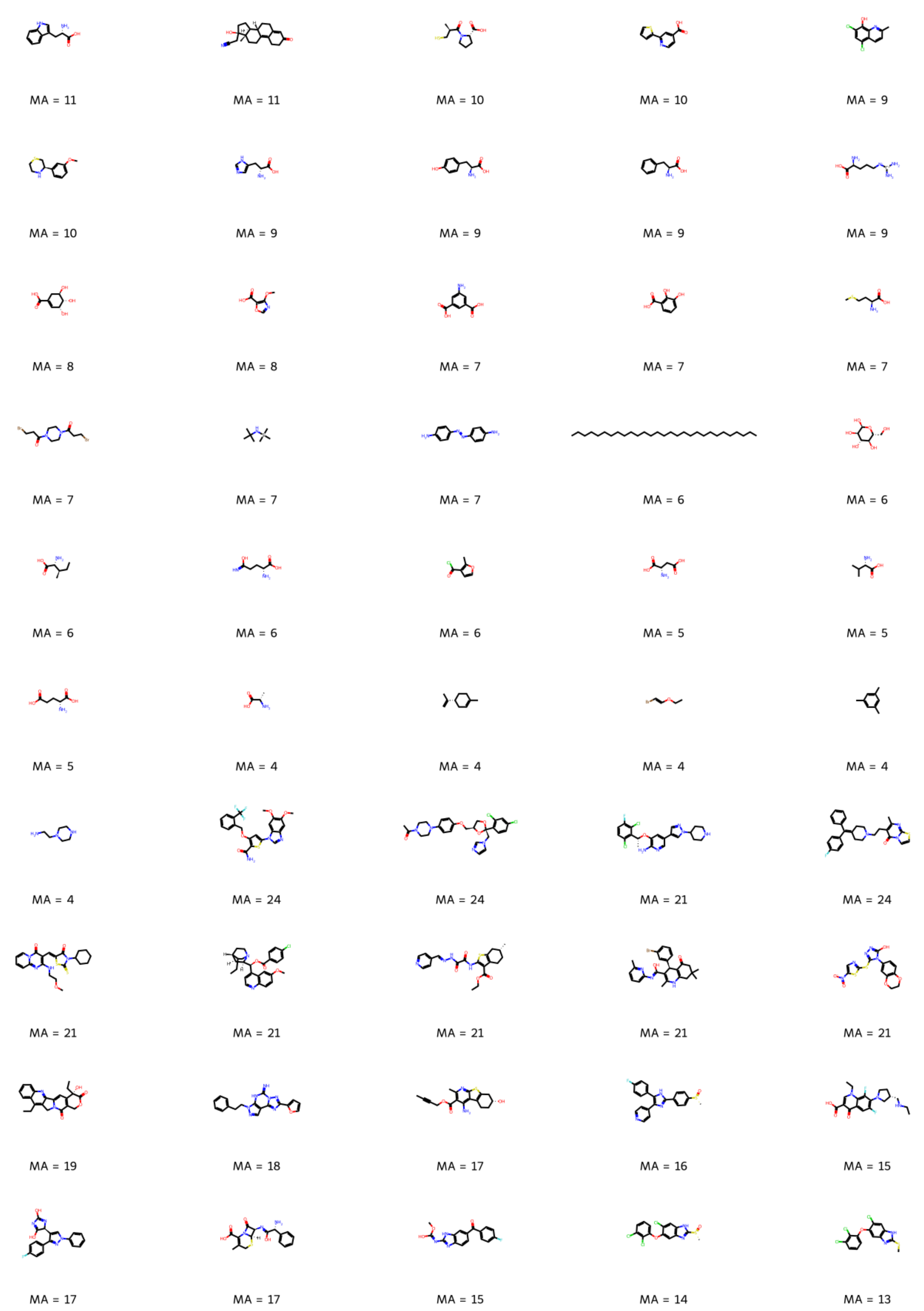


Fig. S19. Structures with calculated molecular assembly (MA) used in the DFT calculated IR study (**Part 2**)

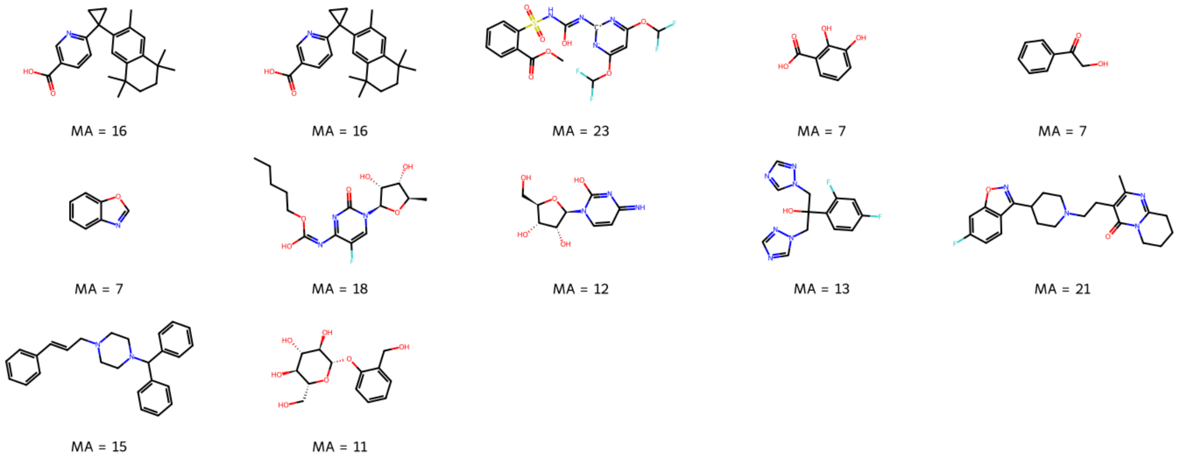


Fig. S20. Structures with calculated molecular assembly (MA) used in the DFT calculated IR study (**Part 3**)

3 Experimental Infrared Spectroscopy

All experimental IR spectra were acquired on a Thermo Scientific Nicolet iS5 with Specac Golden Gate Reflection Diamond ATR System. All samples were measured in their native state at room temperature (solid state unless liquid at room temperature). The acquired data were processed with Thermo Scientific OMNIC 8.3.103 software; using diamond attenuated total reflectance IR (64 scans, resolution 2 cm^{-1}). The spectra were processed at 50% sensitivity and 80% threshold for selecting peaks using OMNIC software (see an example of an acquired spectrum in **Fig. S21**). IR peaks in the fingerprint region (400-1500 cm^{-1}) were counted and correlated against the MA of the molecule. To reduce the error between sample screenings, the background IR spectra were recorded after every 3$^{\text{rd}}$ sample measurement. Linear regression fit between the experimental IR peaks number in the fingerprint region *vs.* MA agreed provided simple model (**Eq. 4**) with a Pearson's correlation coefficient 0.75:

$$\mathrm{MA} = 0.45 \times n_{\mathrm{IR\_peaks}} + 2.26 \qquad \textbf{(4)}$$

where $n_{\mathrm{IR\_peaks}}$ is the number of IR peaks in the region of 400–1500 cm^{-1}. The distribution of MA *vs.* IR-predicted MA is visualised as a histogram in **Fig. S22**. Structures of all compounds used in the study are shown in **Fig. S23** and **Fig. S24**.

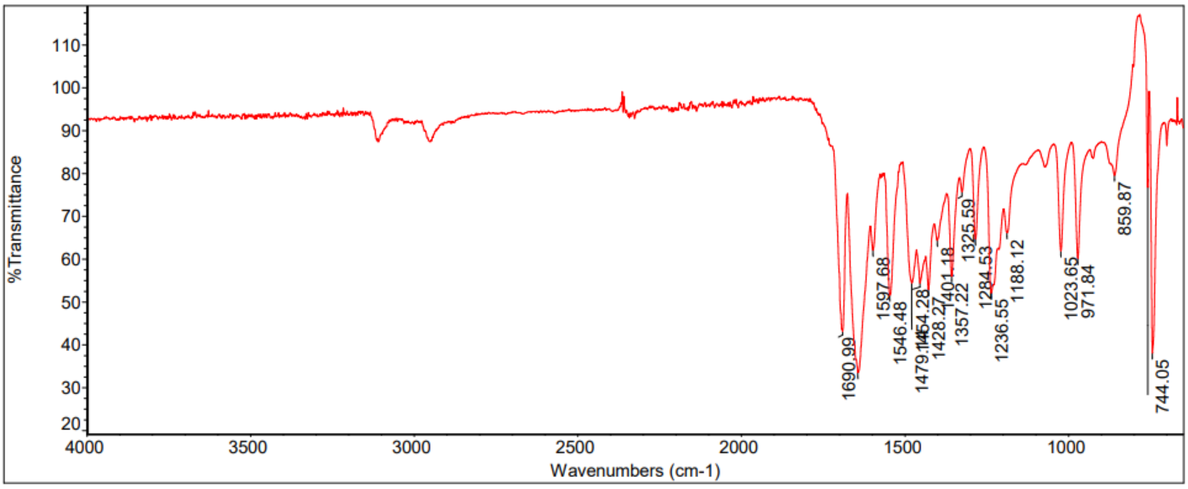


Fig. S21. Example of an experimental IR spectrum of Caffeine. Identified peaks in the fingerprint region 400–1500 cm^{-1} considered in the peak count.

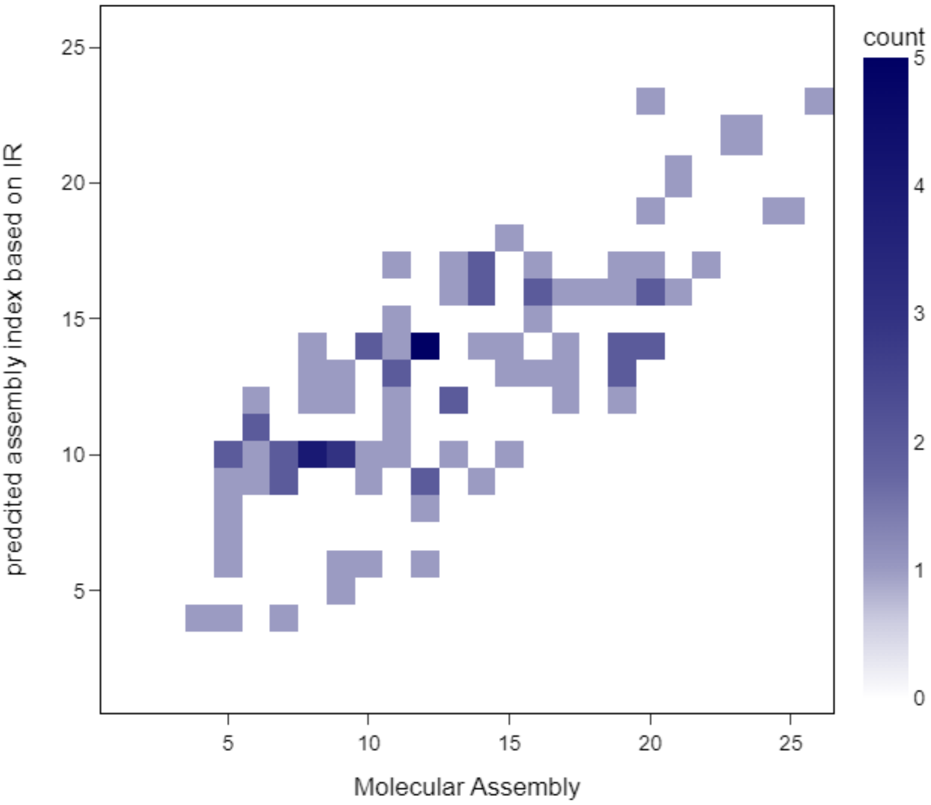


Fig. S22. Histogram of predicted MA based on the 99 experimental IR data using **Eq. 4** vs MA.

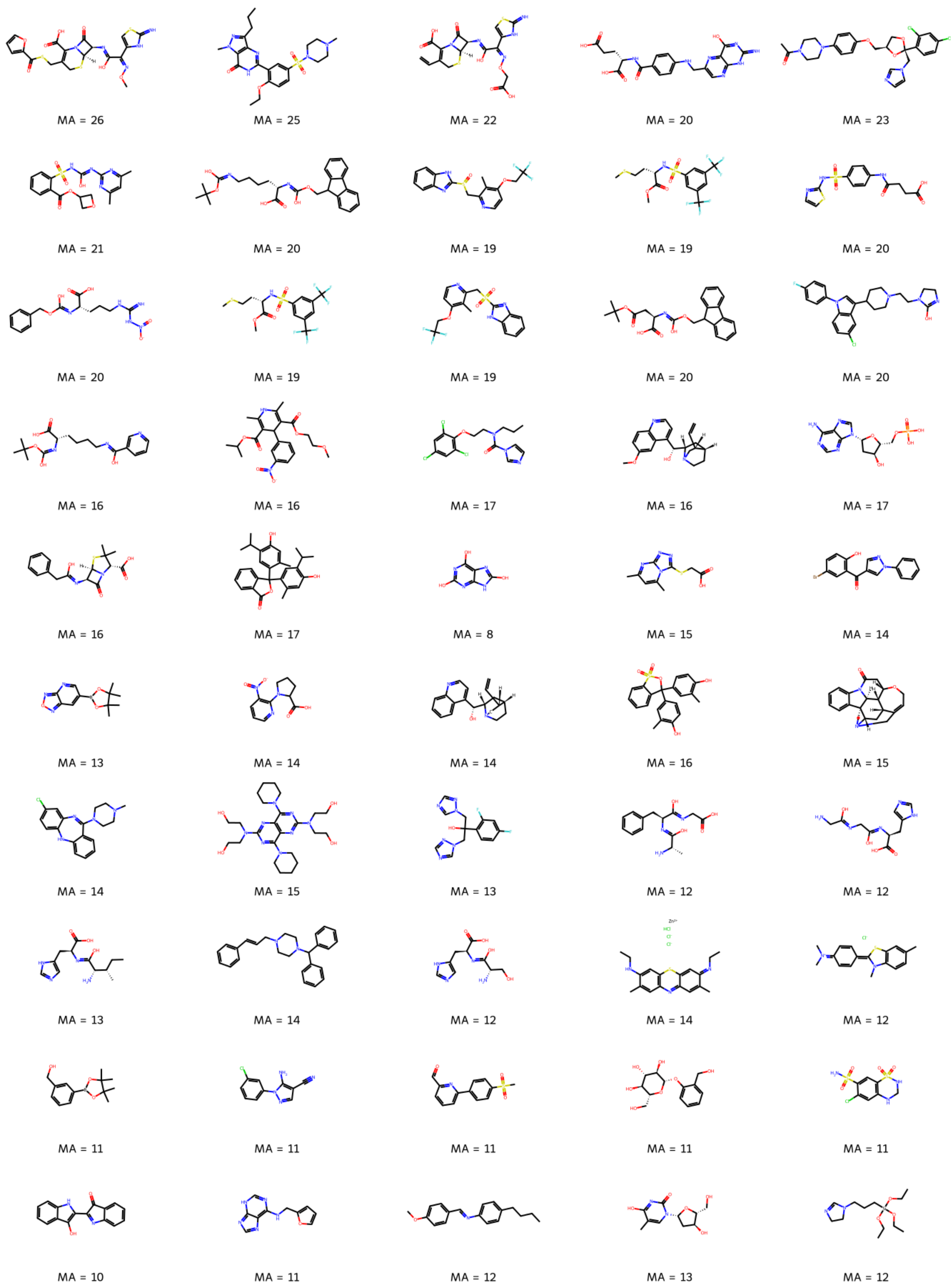
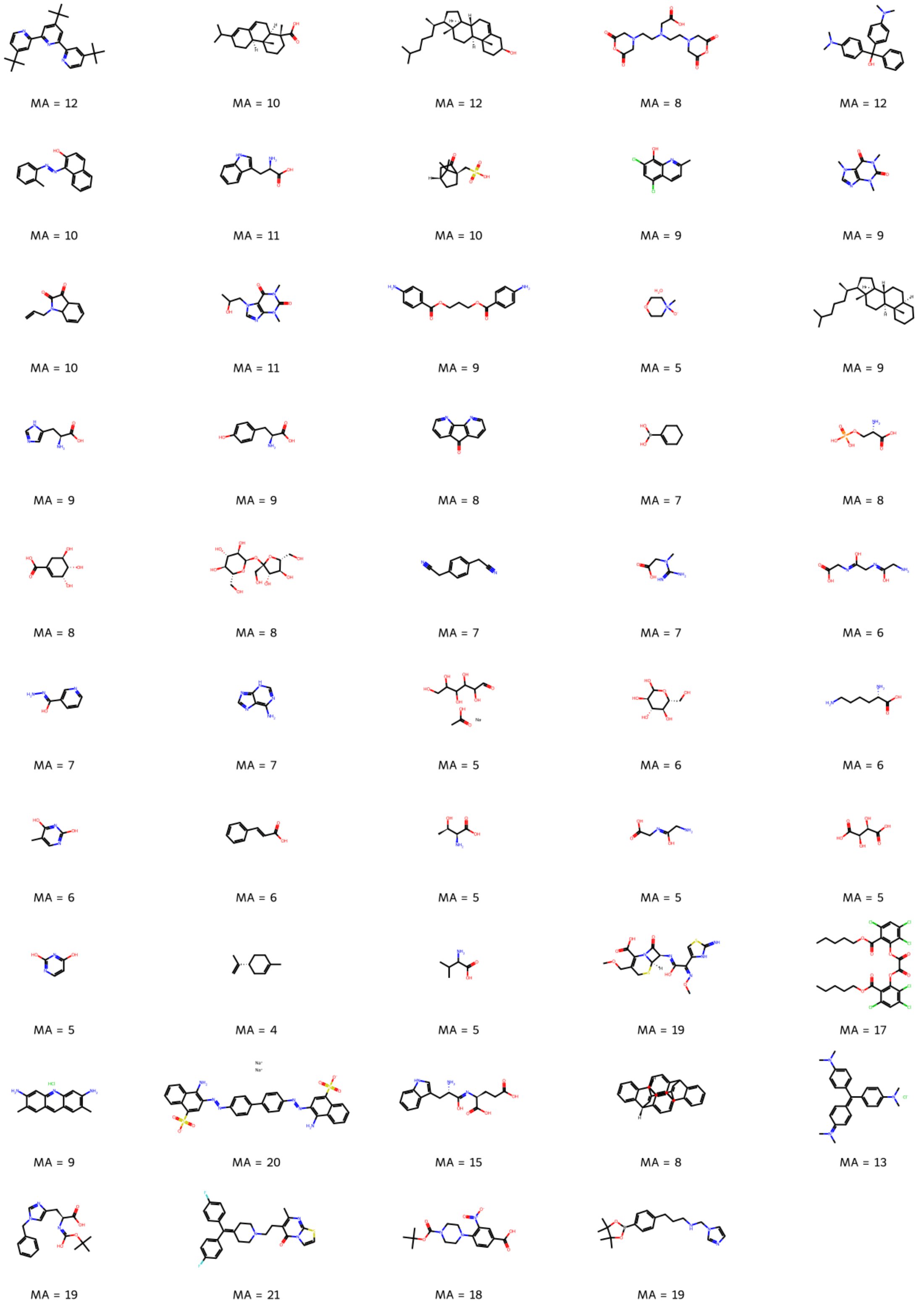


Fig. S23. Structures with calculated molecular assembly (MA) used in the experimental IR study (**Part 1**).



MA = 12 MA = 10 MA = 12 MA = 8 MA = 12

MA = 10 MA = 11 MA = 10 MA = 9 MA = 9

MA = 10 MA = 11 MA = 9 MA = 5 MA = 9

MA = 9 MA = 9 MA = 8 MA = 7 MA = 8

MA = 8 MA = 8 MA = 7 MA = 7 MA = 6

MA = 7 MA = 7 MA = 5 MA = 6 MA = 6

MA = 6 MA = 6 MA = 5 MA = 5 MA = 5

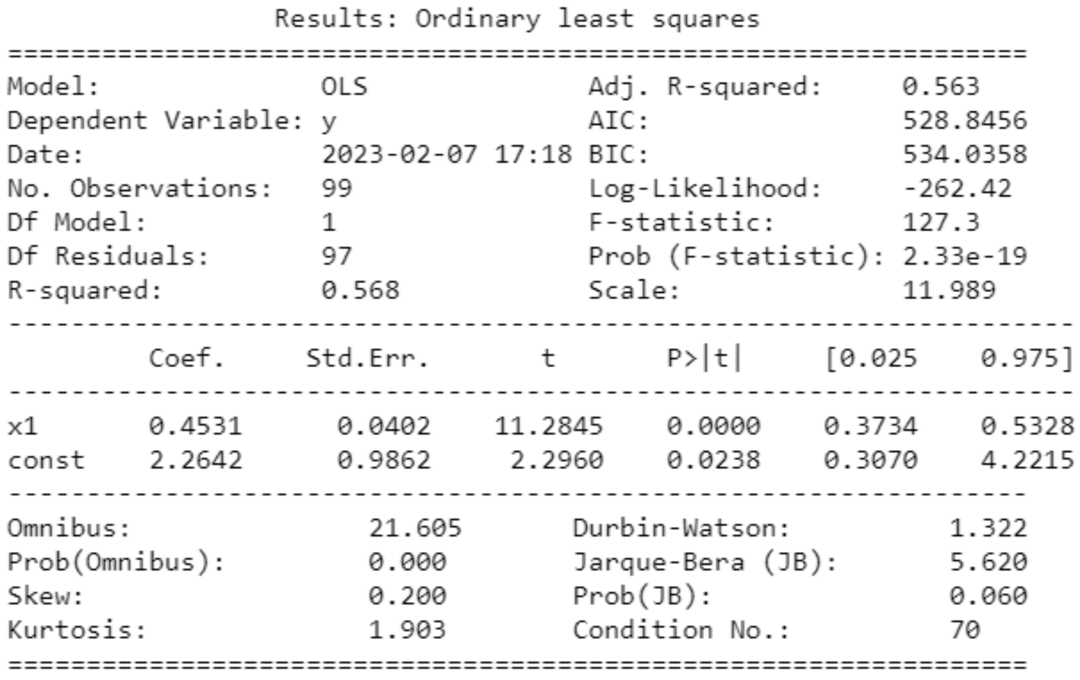
MA = 5 MA = 4 MA = 5 MA = 19 MA = 17

MA = 9 MA = 20 MA = 15 MA = 8 MA = 13

MA = 19 MA = 21 MA = 18 MA = 19

Fig. S24. Structures with calculated molecular assembly (MA) used in the experimental IR study (**Part 2**).

The coefficients for the simple linear function of the number of IR peaks were fi using the *statsmodels.api.OLS* module in python was used (**Fig. S25**).[8]



```
                    Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.563
Dependent Variable: y                AIC:                528.8456
Date:               2023-02-07 17:18 BIC:                534.0358
No. Observations:   99               Log-Likelihood:     -262.42
Df Model:           1                F-statistic:        127.3
Df Residuals:       97               Prob (F-statistic): 2.33e-19
R-squared:          0.568            Scale:              11.989
--------------------------------------------------------------------
         Coef.    Std.Err.       t       P>|t|     [0.025    0.975]
--------------------------------------------------------------------
x1       0.4531     0.0402    11.2845    0.0000    0.3734    0.5328
const    2.2642     0.9862     2.2960    0.0238    0.3070    4.2215
-----------------------------------------------------------------
Omnibus:              21.605       Durbin-Watson:          1.322
Prob(Omnibus):        0.000        Jarque-Bera (JB):       5.620
Skew:                 0.200        Prob(JB):               0.060
Kurtosis:             1.903        Condition No.:          70
=================================================================
```

Fig. S25. Print output from the fit of MA = $x_1 \times n_{\mathrm{peaks}}$ + *const.* for experimental IR spectra; using *statsmodels.api.OLS* in python.[8]

4 Experimental NMR

The investigation of NMR spectroscopy as a prediction tool for MA was experimentally examined on 101 molecules with a range of MA 3–26. Here, the same model that was used to in prediction of MA from the number of different ^{13}C resonances in theoretical dataset was used for
used to predict MA from was used as developed in the theoretical NMR set (**Eq. 1**).

4.1 Sample Preparation details

All samples were prepared with 600 mL d_6-DMSO in 5 mm Bruker 600 MHz rated NMR tube. Concentrations varied from 0.19 mM to 1.17 M due to solubility factors. Several samples (5-aminioisophthalic acid, Oxacillin Sodium Salt, Sildenafil, Triclabendazole) were analyzed *via* NMR at 5, 30 and 300 mM and have shown minimal effect of the concentration on extracted number of chemical environment values from spectra. Samples that did not dissolve in d_6-DMSO were dissolved in a D_2O:d_3-MeCN mixture (ratio 75:25).

4.2 NMR Experimental Parameters:

The NMR used was a Bruker Ascend Aeon 600 MHz NMR spectrometer with a CP DCH 600S3 C/H-D-05 Z cryoprobe installed. All data was processed using Bruker Topspin 3.6.2 and Mestrenova 14.1.1-2451.The NMR experimental parameters are as follows: ^{1}H NMR(16 scans, 20 ppm spectral width, 3.46 second acquisition time, 2.00 second relaxation delay), ^{13}C NMR(128 scans, 250 ppm spectral width, 1.73 second acquisition time, 0.80 second relaxation delay), ^{13}C DEPTQ 90 and DEPTQ 135 (64 scans,250 ppm spectral width,1.73 seconds acquisition time, 1.00 second relaxation delay), ^{1}H PSYCHE(16 scans, 12.49 ppm spectral width, 0.89 seconds acquisition time, 1.00 second relaxation delay), HSQC (8 scans, 10 ppm spectral width F2, 250 spectral width F1, 0.09 second acquisition time, 1.49 seconds relaxation delay), ^{13}C DOSY (pseudo 2D experiment scans, 200 ppm spectral width F2, 8 TD points, 1.10 second acquisition time, 8.00 second relax delay, 0.80 second diffusion time d20, 1450 μsec gradient pulse P30).

The experiments were tested on several examples to cover range of concentrations from 5 mM to 300 mM to prove the same number of individual carbon peak types can be achieved, regardless the concentration.

4.3 Classification of the Carbon Types

Accurate counts for all individual ^{13}C type environments were obtained using a combination of ^{13}C, DEPTQ135, DEPTQ90 and HSQC analysis to ensure maximum accuracy before unblinding samples.

To determine degree of substitution for ^{13}C chemical environments, this was first approached using two types of ^{13}C DEPT experiment, DEPTQ 135 and DEPTQ 90, that phase carbon signals positive or negative depending on their degree of substitution. The DEPTQ experiment was chosen (not the DEPT) to detected quaternary carbons (otherwise not detected by DEPT). The 135 and 90 portion stands for the final ^{1}H tip angle of the pulse in the pulse program before acquisition.[19,20]

First, ^{13}C DEPTQ 135 observes quaternary and CH_2 peaks as one phase and the CH and CH_3 peaks as the other. The ^{13}C DEPTQ 90 then is measured to compliment the ^{13}C DEPTQ 135 with only detection of quaternary peaks in one phase and the CH peaks in the other. Using DEPTQ135 and DEPTQ 90 together, all degrees of substitution of the carbons can be identified via NMR. The solvent (expected as quaternary if deuterated) peak must be disregarded in the two counts of carbon peaks. To verify assignment of degree of unsubsitution for ^{13}C chemical environments in blind samples, ^{1}H-^{13}C HSQC was used. In the HSQC experiment, the peaks are phased as they are in DEPT with CH_2 cross peaks in one phase and CH and CH_3 in the other. As CH_2 cross peaks are detected in their phase alone, this provides an easy method of counting number of CH_2 chemical environments from CH_2 cross peaks. Quaternary ^{13}C cross peaks are not detected in HSQC.

The herein described workflow is illustrated on quinine, and its ^{13}C NMR (**Fig. S26**), DEPTQ-90 (**Fig. S27**), DEPTQ-135 (**Fig. S28**) and ^{1}H-^{13}C HSQC (**Fig. S29**) spectra.
Structures of all compounds used in the NMR study are in **Fig. S30** and **Fig. S31**.

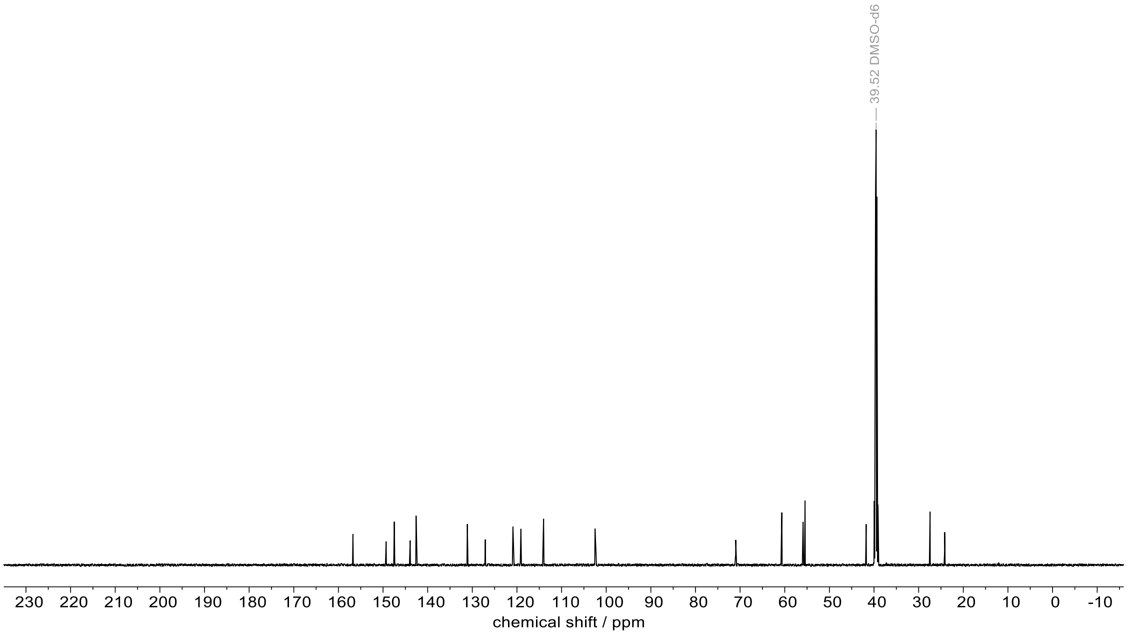


Fig. S26. ^{13}C NMR (150 MHz) of quinine.

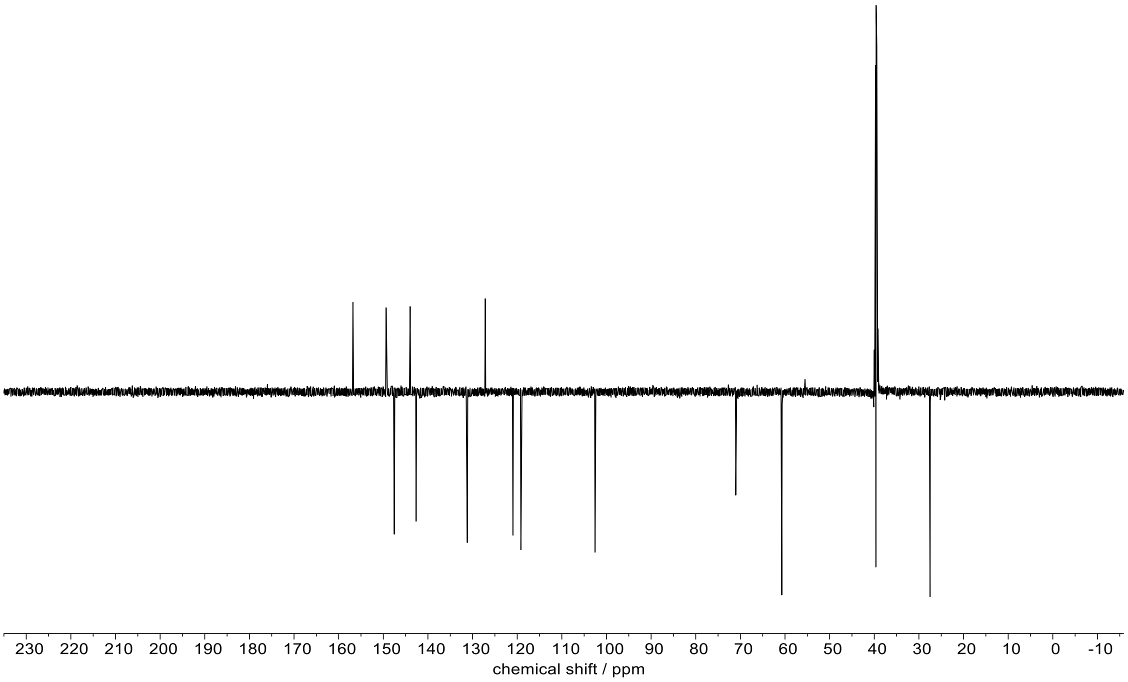


Fig. S27. DEPTQ-90 ^{13}C NMR (150 MHz) of quinine. Peaks at positive phase are C, peaks at negative are CH. The solvent peak needs to be subtracted from the C count (in this case DMSO-d6 in the positive phase).

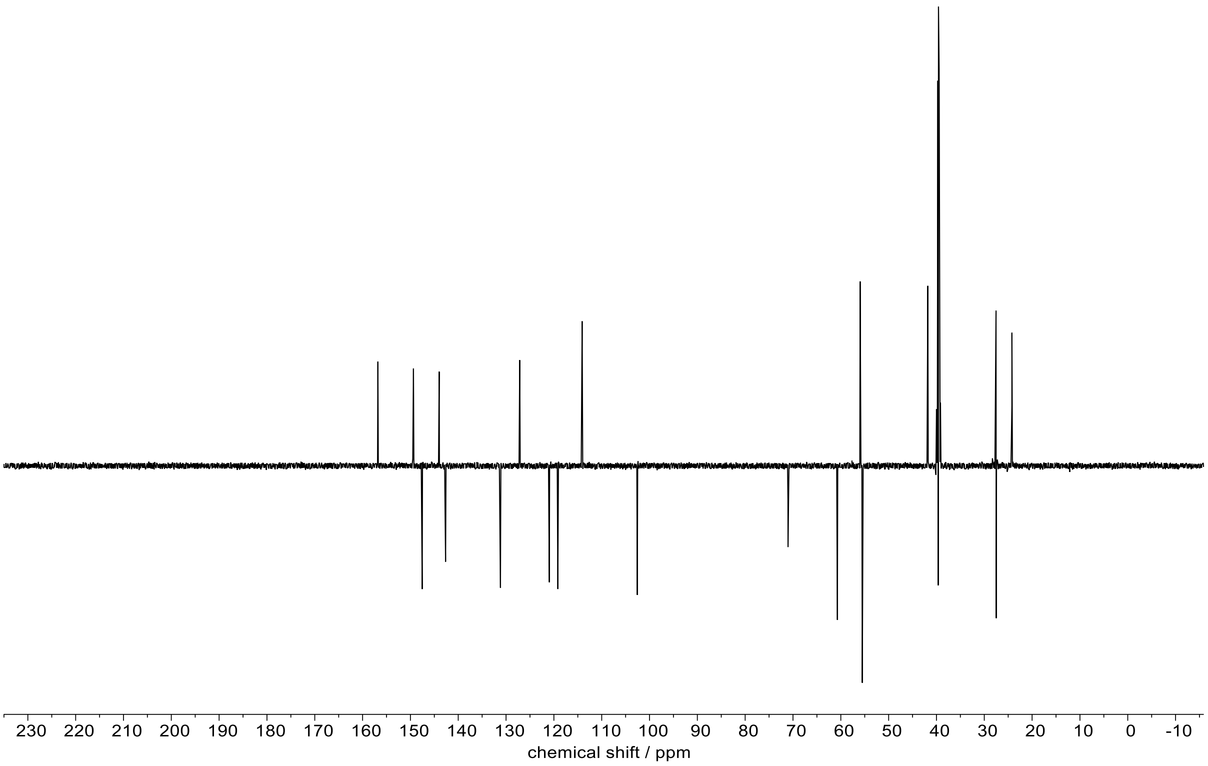


Fig. S28. DEPTQ-135 ^{13}C NMR (150 MHz) of quinine. Peaks at positive phase are C and CH_2, peaks at negative as CH and CH_3.

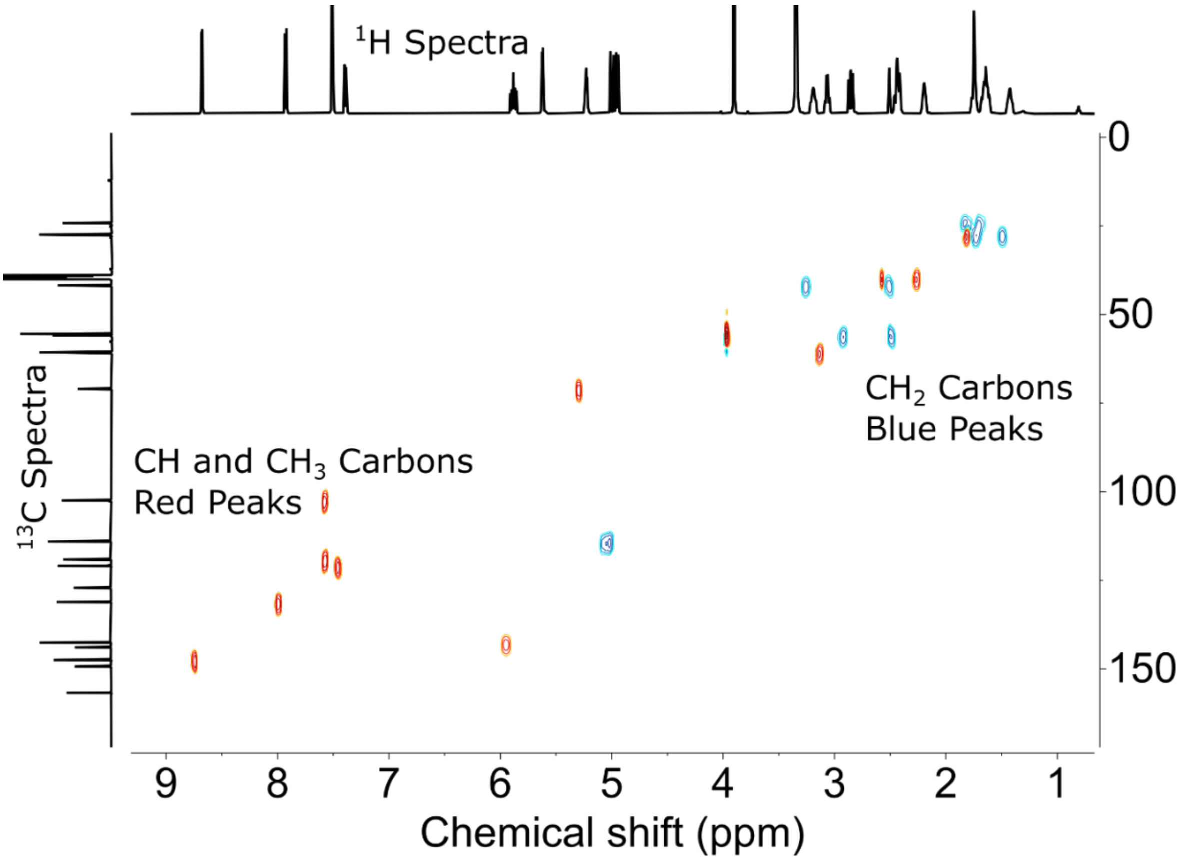


Fig. S29. ^1H-^{13}C HSQC Spectra of Quinine- Crosspeaks in red note CH and CH_3, blue note CH_2

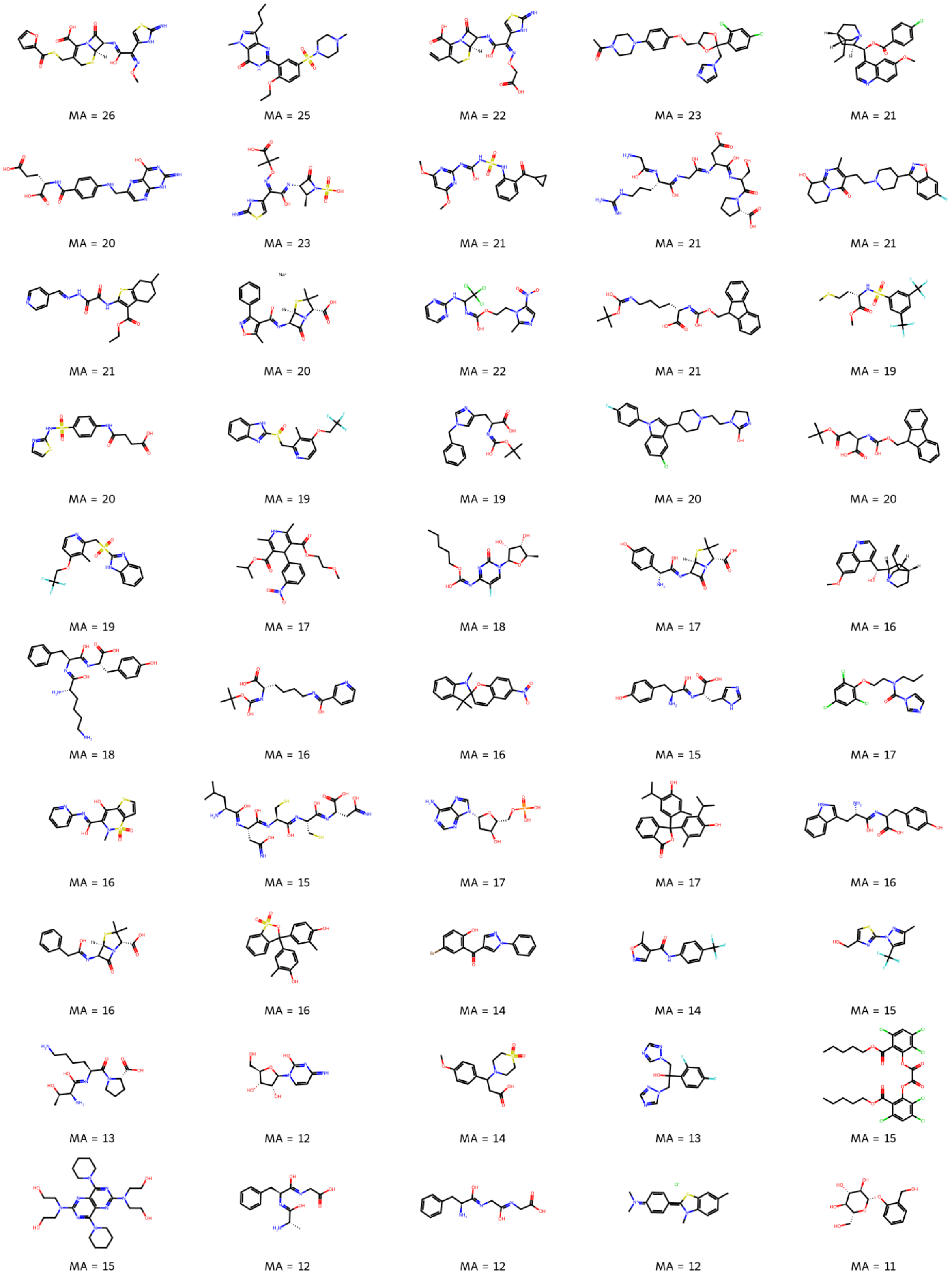


Fig. S30. Structures with calculated molecular assembly (MA) used in the experimental NMR study (part 1).

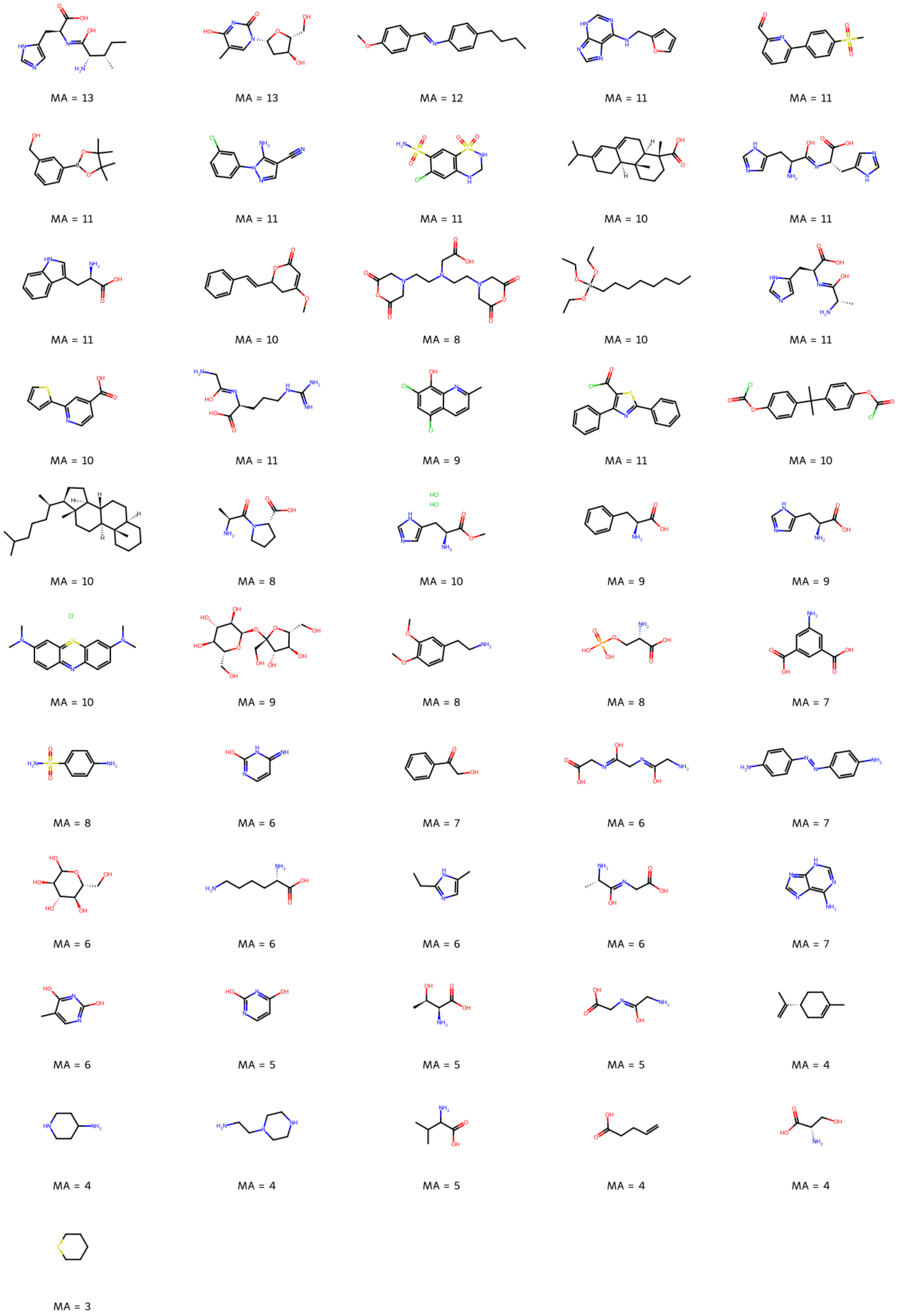


Fig. S31. Structures with calculated molecular assembly (MA) used in the experimental NMR study (part 2).

5 Combing IR and NMR data

On the set of 10,000 calculated NMR and IR spectra, we have examined our hypothesis that combined information can provide a more reliable MA prediction. We have used the models for the individual spectroscopic techniques (**Eq. 1** and **Eq. 2**) and allowed them to optimise for their relative weighting. The combined model provided a higher correlation of 0.90 using the weighted average of 0.55×NMR and 0.45×IR inferred MA (**Fig. S32**).

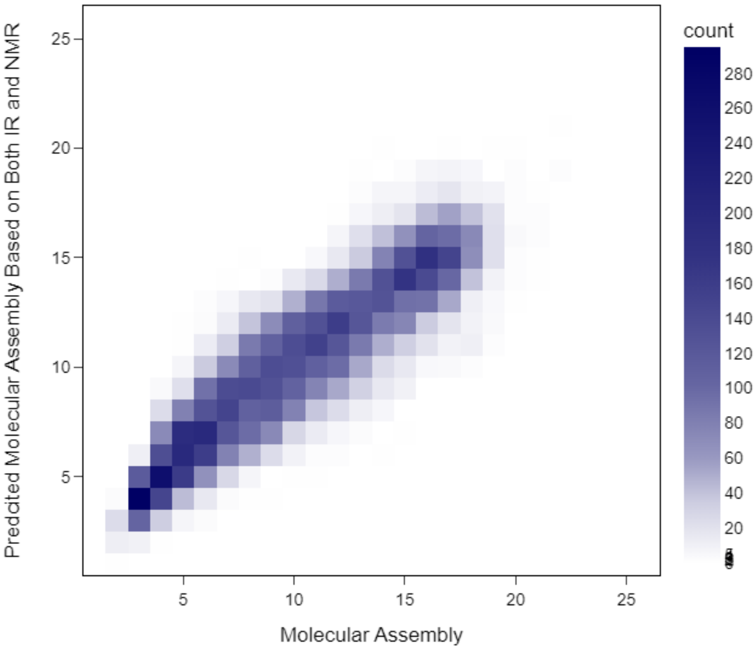


Fig. S32. Histogram of predicted MA vs expected MA on 10,000 compounds sample based on the fit of NMR and IR prediction using **Eq. 1** and **Eq. 2**, respectively, using weighted average of NMR and IR of 0.55 and 0.45, respectively.

Analogously to the simulated data, the ratio for weighted average of the models based on the **Eq. 1** for NMR and the experimental model fit for IR **Eq. 4** were optimised for the experimental test sample on the available intersection of the experimental NMR and IR data, comprising 55 molecules. The weighting was 0.7 and 0.3 for ratio of NMR and IR MA predictions, respectively, yielding correlation of the predicted and experimental MA of 0.89 (**Fig. S33**).

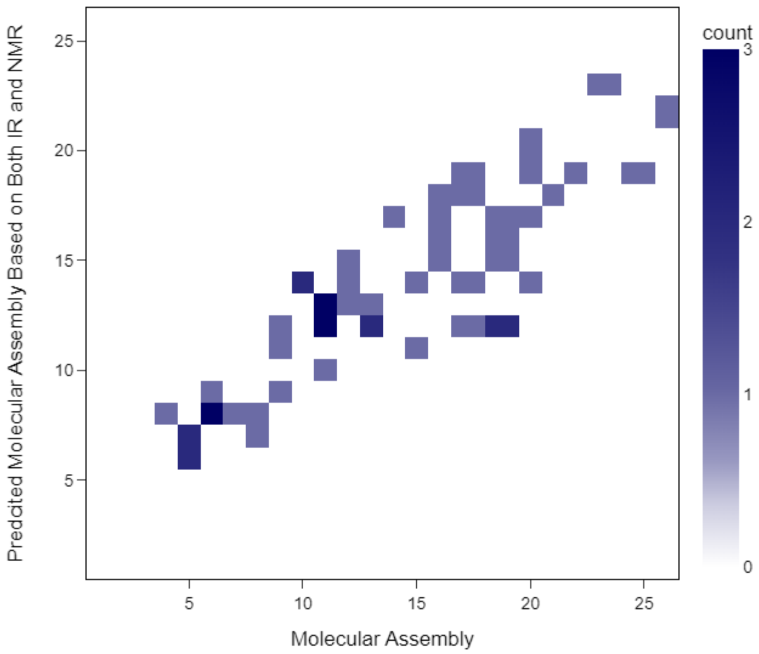


Fig. S33. Histogram of predicted MA vs expected MA on 55 compounds sample based on the fit of NMR and IR prediction using **Eq. 1** and **Eq. 4**, respectively, using weighted average of NMR and IR of 0.7 and 0.3, respectively.

6 Mixture Analysis

6.1 NMR

To deconvolute the mixture *via* NMR as a proof of concept, the mixture of two compounds was examined using ^{13}C DOSY. The experimental setup of the ^{13}C DOSY was 200 ppm spectral width, 8 TD points, 1.10 second acquisition time, 8.00 second relax delay, 0.80 second diffusion time d20, 1450 µsec gradient pulse P30 using the ^{13}C DOSY-stebpgppg1s routine with 256 scans. An example of quinine and 5-aminoisopthalic acid is presented in **Fig. S34**.

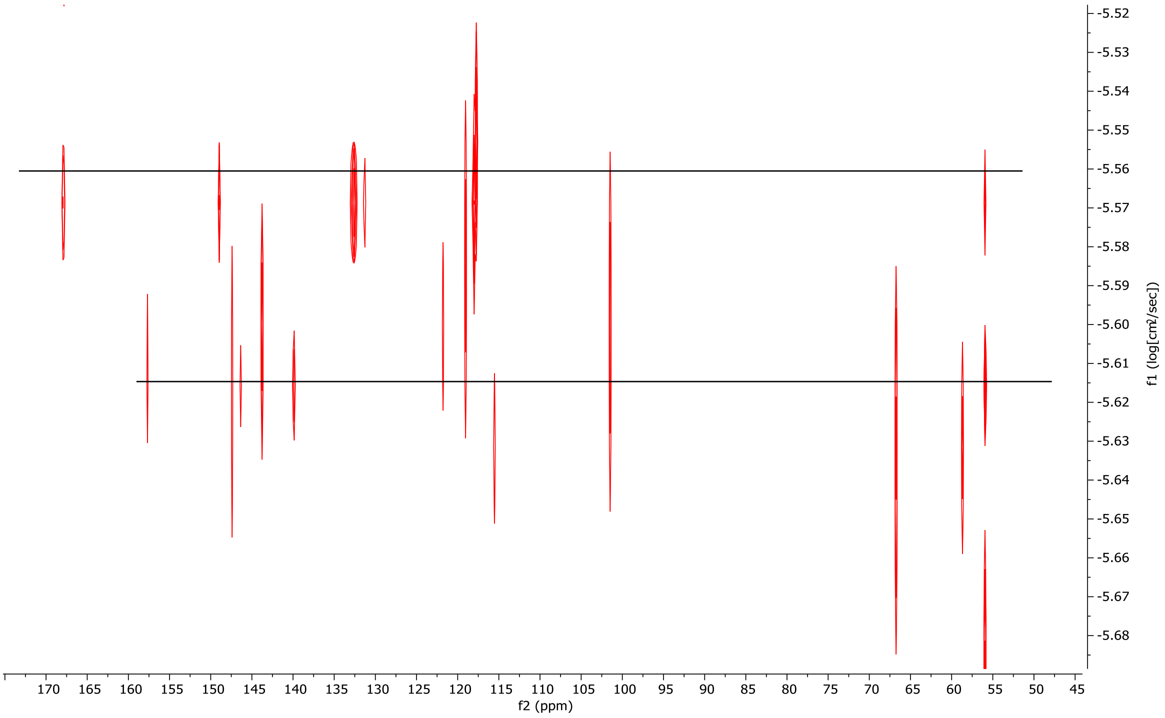
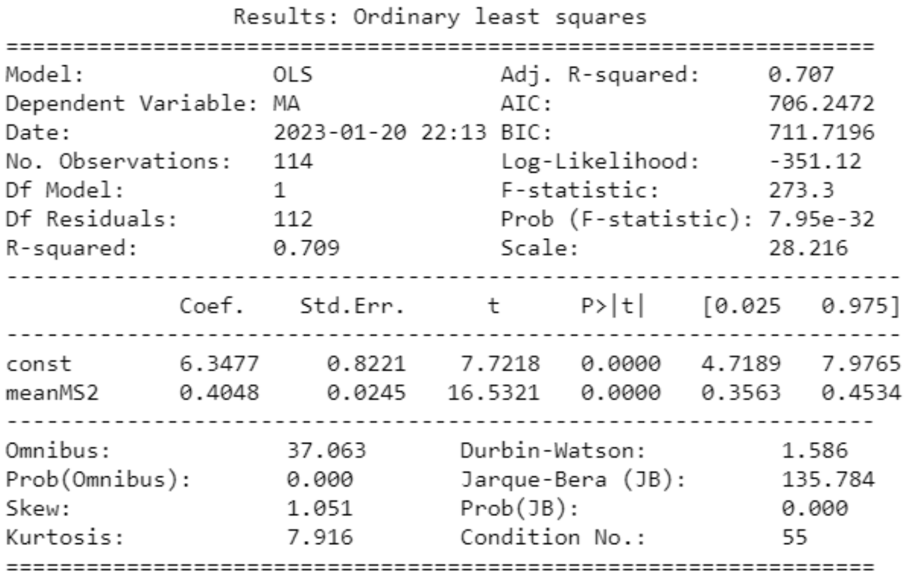


Fig. S34. ^{13}C DOSY of quinine and 5-aminoisophthalic acid mixture. Two horizontal lines guides the separation of the ^{13}C signals.

6.2 LC-MS

MA was measured by Mass Spectrometry by counting MS peaks after MS^2 fragmentation with exact post processing as described previously by our group in Nature Communications.[1] Individual sample underwent a 15 µl injection from a Advion Nanomate followed by a MS^1 full scan and MS^2 fragmentation using a Thermo Fusion Lumos Orbitrap. The mixture was analysed by chromatographic separation with a 15cm C18 column (Agilent) on a Thermo U300 HPLC system with a flow rate of 0.4 ml/min for 35 minutes using a 10 µL injection. The test sample was equimolar mixture of 10 components in 80% MeOH in H_2O. The resolution for MS^1 and MS^2 was set at 240000 and 30000, respectively.

Compound	MS2 Peak Count	Inferred MA	MA
Ceftiofur	31	19	26
Sildenafil	37	21	25
Ketoconazole	47	25	24
Folic Acid	43	24	20
Succinylsulfathiazole	44	24	20
Sucrose	n.a.	—	8
Glucose	n.a.	—	6
Uracil	6	9	5
L-Valine	3	8	5
S-Limonene	n.a.	—	4



```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.707
Dependent Variable: MA               AIC:                706.2472
Date:               2023-01-20 22:13 BIC:                711.7196
No. Observations:   114              Log-Likelihood:     -351.12
Df Model:           1                F-statistic:        273.3
Df Residuals:       112              Prob (F-statistic): 7.95e-32
R-squared:          0.709            Scale:              28.216
--------------------------------------------------------------------
             Coef.    Std.Err.      t      P>|t|    [0.025   0.975]
--------------------------------------------------------------------
const        6.3477     0.8221    7.7218   0.0000   4.7189   7.9765
meanMS2      0.4048     0.0245   16.5321   0.0000   0.3563   0.4534
------------------------------------------------------------------
Omnibus:             37.063       Durbin-Watson:          1.586
Prob(Omnibus):       0.000        Jarque-Bera (JB):       135.784
Skew:                1.051        Prob(JB):               0.000
Kurtosis:            7.916        Condition No.:          55
==================================================================
```

Fig. S35. Print output from the fit of MA = x_1×n_{peaks} + *const.* for experimental MS spectra; using *statsmodels.api.OLS* in python.[8]

7 References

1. Marshall, S. M. *et al.* Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nat Commun* **12**, 3033 (2021).
2. Rücker, G. & Rücker, C. Automatic Enumeration of All Connected Subgraphs. *MATCH Comm. in Math. in Comp. Chem.* 145–149.
3. McKay, B. D. & Piperno, A. Practical graph isomorphism, II. *Journal of Symbolic Computation* **60**, 94–112 (2014).
4. Liu, Y. *et al.* Exploring and mapping chemical space with molecular assembly trees. *Sci. Adv.* **7**, eabj2465 (2021).
5. Marshall, S. M., Moore, D. G., Murray, A. R. G., Walker, S. I. & Cronin, L. Formalising the Pathways to Life Using Assembly Spaces. *Entropy* **24**, 884 (2022).
6. 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: Lab administration, spectra assignment aid and local database. *Magn. Reson. Chem.* **53**, 582–589 (2015).
7. Landrum, G. *et al.* rdkit/rdkit: 2022_09_4 (Q3 2022) Release. (2023) doi:10.5281/ZENODO.591637.
8. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in 92–96 (2010). doi:10.25080/Majora-92bf1922-011.
9. Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. *WIREs Comput Mol Sci* **11**, (2021).
10. Jablonka, K. M., Patiny, L. & Smit, B. Making Molecules Vibrate: Interactive Web Environment for the Teaching of Infrared Spectroscopy. *J. Chem. Educ.* **99**, 561–569 (2022).
11. Python API for the extended tight binding program. https://github.com/grimme-lab/xtb-python.

12. Padamati, S. K. *et al.* Transient Formation and Reactivity of a High-Valent Nickel(IV) Oxido Complex. *J. Am. Chem. Soc.* **139**, 8718–8724 (2017).

13. te Velde, G. *et al.* Chemistry with ADF. *J. Comput. Chem.* **22**, 931–967 (2001).

14. Swart, M. & Bickelhaupt, F. M. QUILD: QUantum-regions interconnected by local descriptions. *J. Comput. Chem.* **29**, 724–734 (2008).

15. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).

16. Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **33**, 8822–8824 (1986).

17. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **132**, 154104 (2010).

18. Klamt, A. & Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* 799–805 (1993) doi:10.1039/P29930000799.

19. Burger, R. & Bigler, P. DEPTQ: Distorsionless Enhancement by Polarization Transfer Including the Detection of Quaternary Nuclei. *Journal of Magnetic Resonance* **135**, 529–534 (1998).

20. Bigler, P., Kümmerle, R. & Bermel, W. Multiplicity editing including quaternary carbons: improved performance for the13C-DEPTQ pulse sequence. *Magn. Reson. Chem.* **45**, 469–472 (2007).