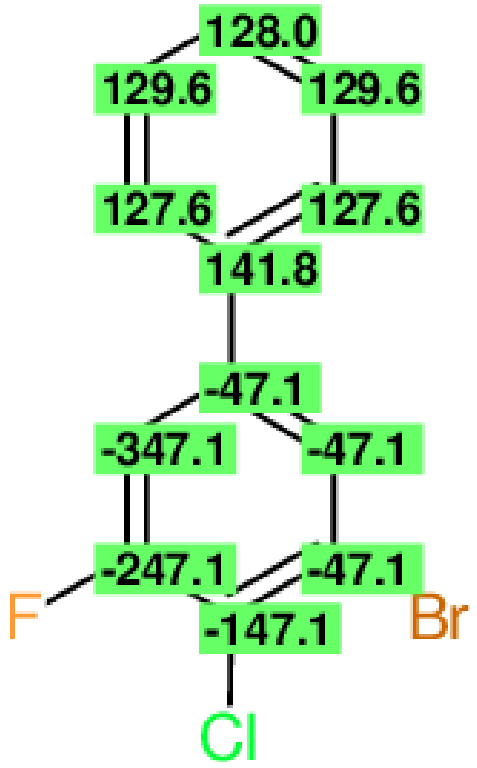
Automatisierte Strukturrevisionen mittels CSEARCH

Wolfgang Robien
Institut für organische Chemie, Universität Wien

23. März 2021 / TU Berlin via Zoom

Peer-Reviewing von Datensätzen – Einfluss von falschen Daten auf die Vorhersage

Diese Verbindung ist nicht bei CAS registriert (Stand: 19. März 2021)



Hier Literaturwerte von Biphenyl genommen

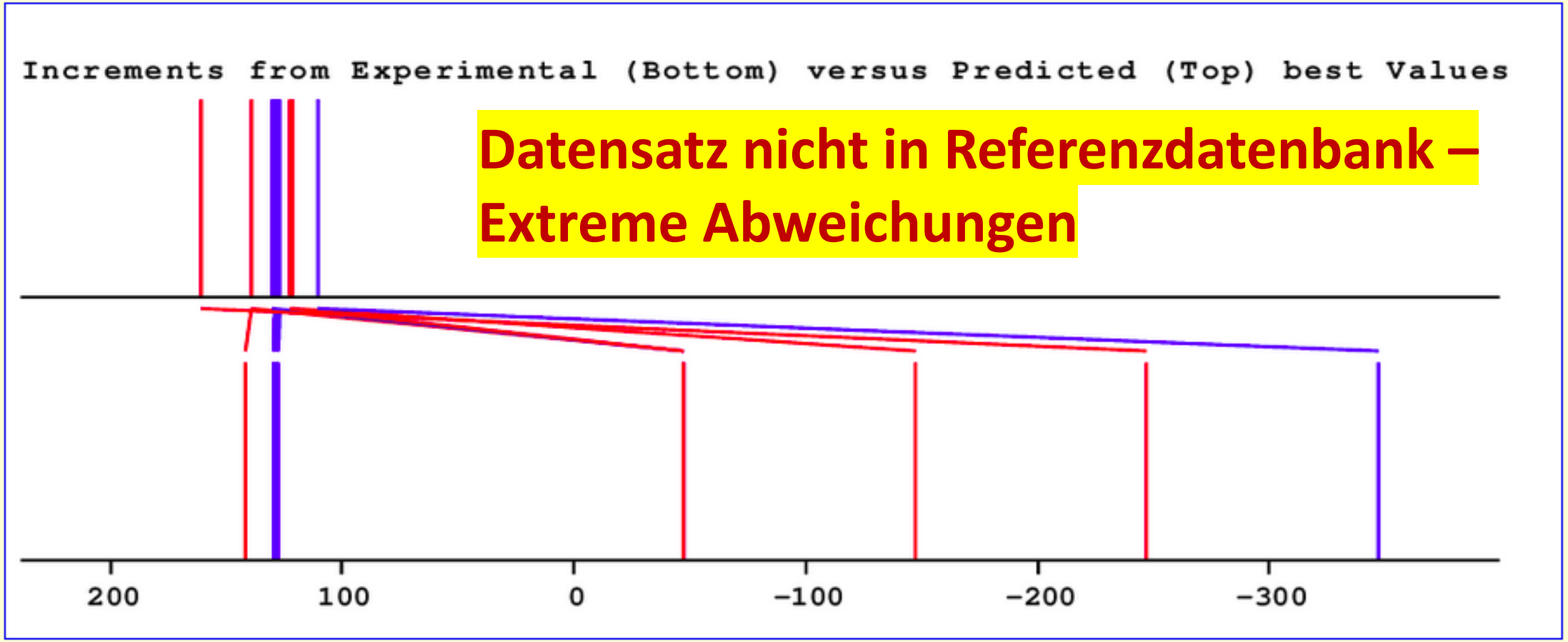
Hier vollkommen sinnbefreite Werte genommen (-47.11 / -147.11 / -247.11 / -347.11)

Datensatz somit extrem falsch + unbrauchbar

Bewertung mittels CSEARCH-Robot-Referee

Falscher Datensatz in der Datenbank nicht vorhanden

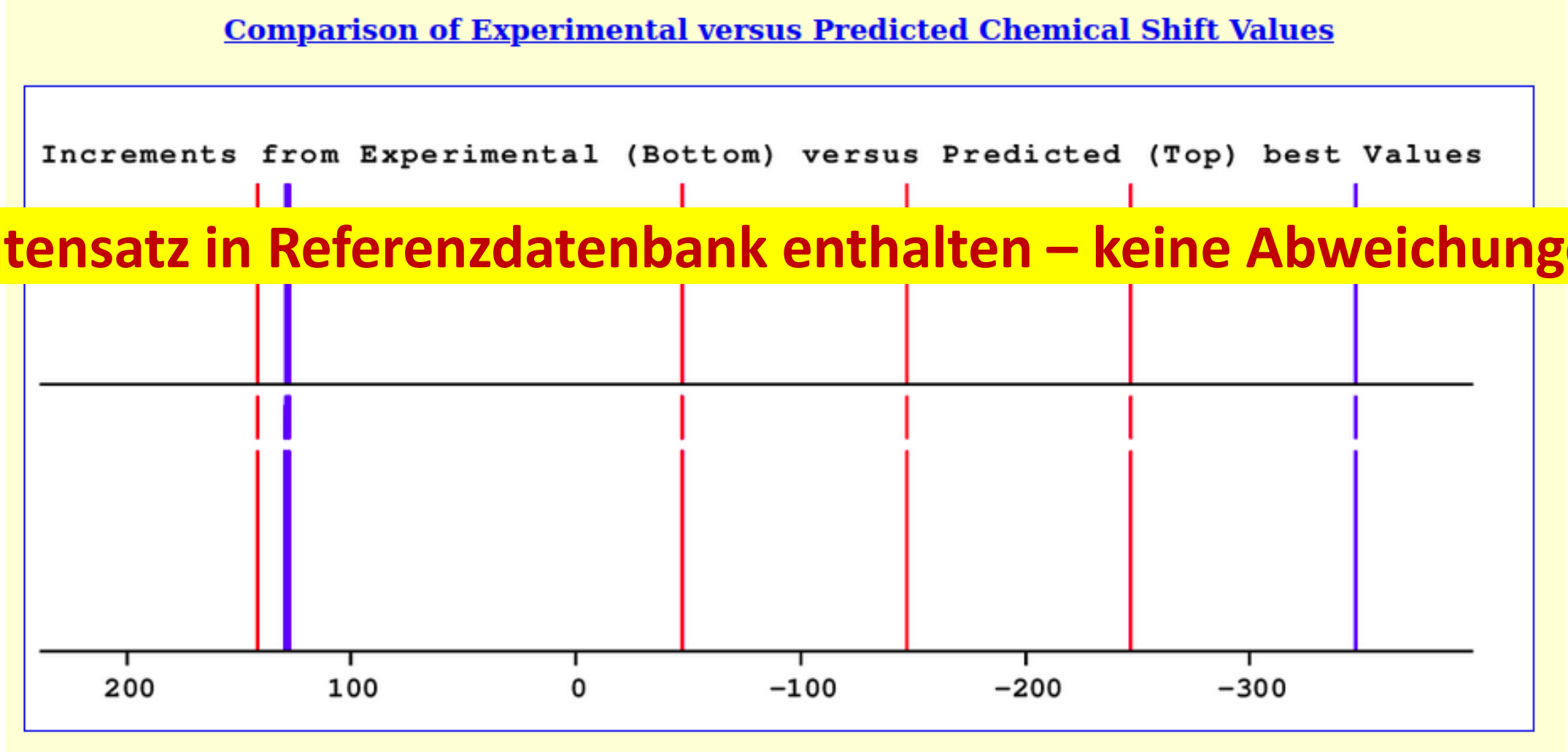
Comparison of Experimental versus Predicted Chemical Shift Values



Datensatz nicht in Referenzdatenbank – Extreme Abweichungen

Bewertung mittels CSEARCH-Robot-Referee

Falscher Datensatz in der Datenbank vorhanden



Datensatz in Referenzdatenbank enthalten – keine Abweichungen

Schlussfolgerungen für Vorhersage mittels HOSE-Code:

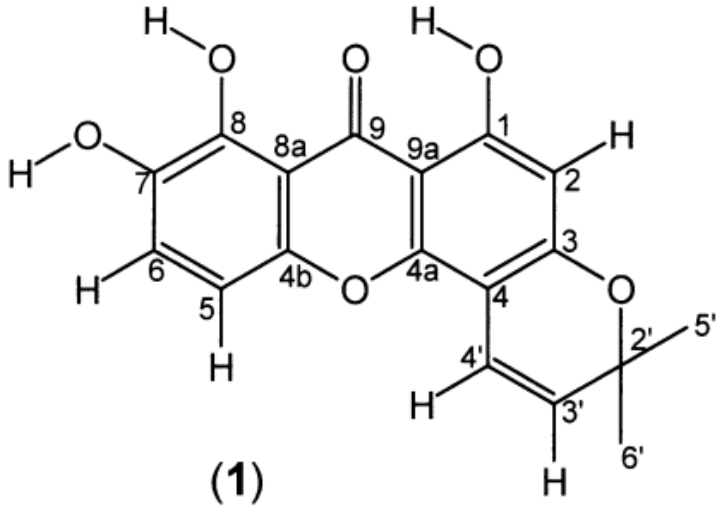
HOSE-Code - von nahezu allen Vorhersageprogrammen verwendet

Reproduziert exakt den Datenbankinhalt aufgrund des verwendeten mathematischen Modells

Verwendete Referenzdaten korrekt → Ergebnis korrekt, weil konsistent
Verwendete Referenzdaten falsch → Ergebnis falsch, weil konsistent

Konsistenz zwischen Datenbankinhalt und Vorhersage immer gegeben
Richtigkeit nur gegeben, wenn dahinterstehende Daten richtig sind

Praxisbeispiel: Globulixanthone C aus Phytochemistry 2002



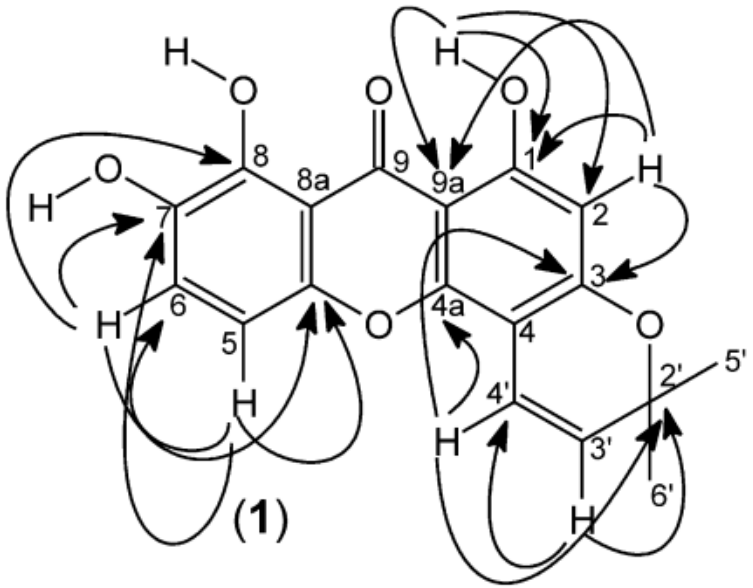


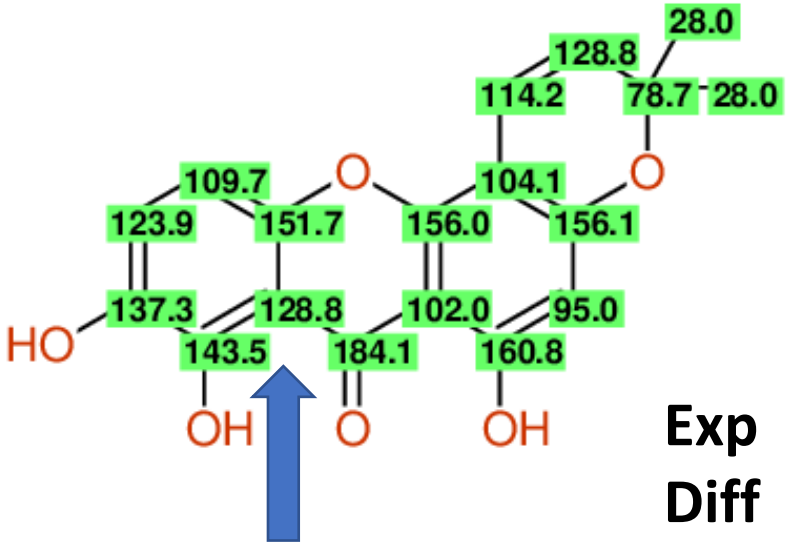
Fig. 1. Significant long-range correlations observed in ^{13}C–^{1}H HMBC for compound **1** in DMSO.

^{1}H (300 MHz) and ^{13}C (75 MHz) assignments

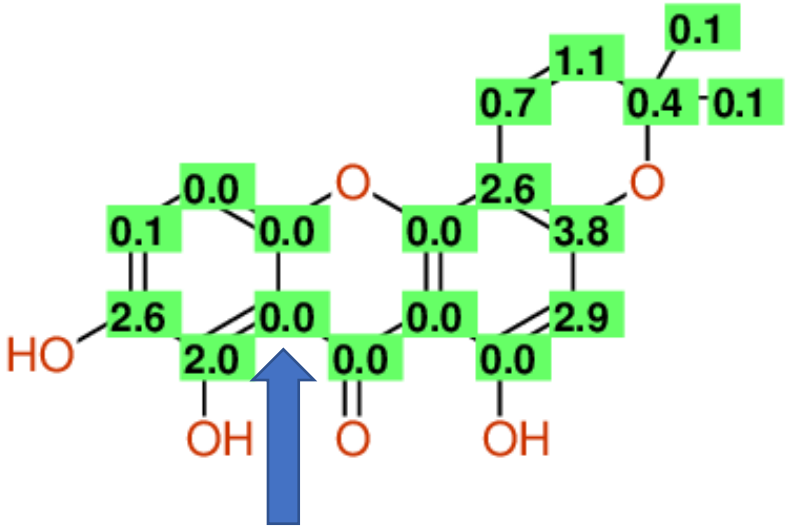
1 (DMSO)

Attribution	^{13}C	^{1}H [m, *J* (Hz)]
1	160.8	–
2	95.0	6.40 (*s*)
3	156.1	
4	104.1	–
4a	156.0	–
4b	151.7	–
5	109.7	6.67 (*d*, 8.1)
6	123.9	7.26 (*d*, 8.1)
7	137.3	–
8	143.5	–
8a	128.8	–
9	184.1	–
9a	102.0	–
2′	78.7	–
3′	128.8	5.85 (*d*, 10.0)
4′	114.2	6.65 (*d*, 10.0)
5′	28.0	1.49 (*s*)
6′	28.0	1.49 (*s*)
1-OH	–	12.85 (*s*)
7-OH	–	8.50 (*s*, *brs*)
8-OH	–	11.10 (*s*)

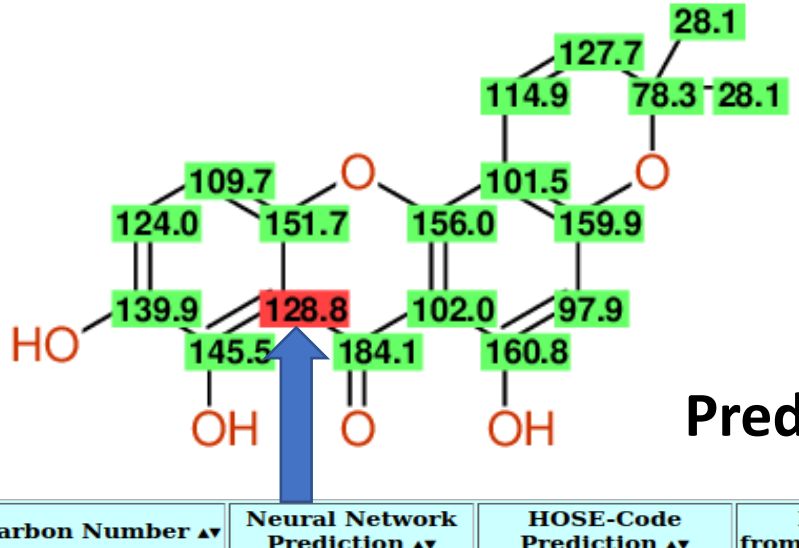
State-of-the art 1D- und 2D-NMR, H1, C13,
HMBC, NOESY zusammen mit IR und MS
Übersichtliche Präsentation im Paper
Datensatz in CSEARCH enthalten



Exp



Diff



Pred

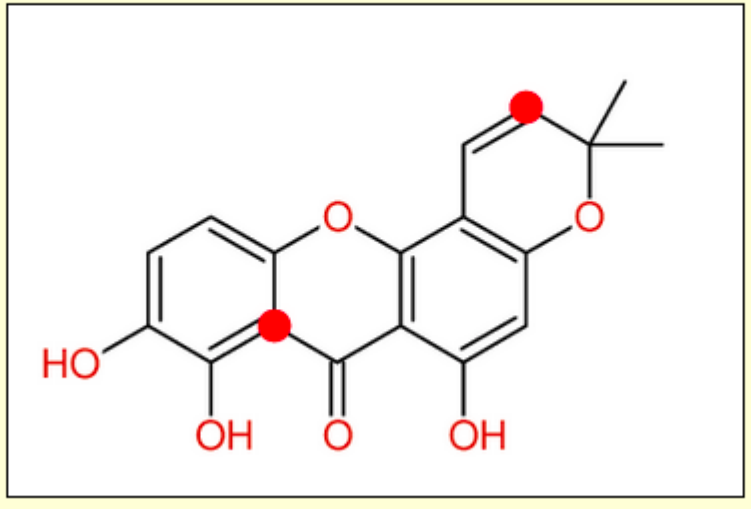
Vorhersagen (HOSE & NN) deutlich unterschiedlich

Carbon Number	Neural Network Prediction	HOSE-Code Prediction	Preferred Value from both Predictions	Experimental values	Difference (Exp-Pred/ppm)
1	161.6	160.8	160.8	160.8	0.0
2	99.8	97.9	97.9	95.0	2.9
3	161.9	159.9	159.9	156.1	3.8
4	99.9	101.5	101.5	104.1	2.6
5	153.5	156.0	156.0	156.0	0.0
6	148.3	151.7	151.7	151.7	0.0
7	108.9	109.7	109.7	109.7	0.0
8	126.4	124.0	124.0	123.9	0.1
9	143.8	139.9	139.9	137.3	2.6
10	148.4	145.5	145.5	143.5	2.0
11	108.6	128.8	128.8	128.8	0.0
12	184.9	184.1	184.1	184.1	0.0
13	100.4	102.0	102.0	102.0	0.0
14	77.2	78.3	78.3	78.7	0.4
15	128.2	127.7	127.7	128.8	1.1
16	120.3	114.9	114.9	114.2	0.7
17	27.8	28.1	28.1	28.0	0.1
18	27.8	28.1	28.1	28.0	0.1

Bewertung ergibt: Minor Revision - Begründung für diese Bewertung:

Große Differenz zwischen HOSE-Code und NN-Vorhersage
Mögliches Symmetrieproblem (Linie bei 128.8ppm 2x verwendet)

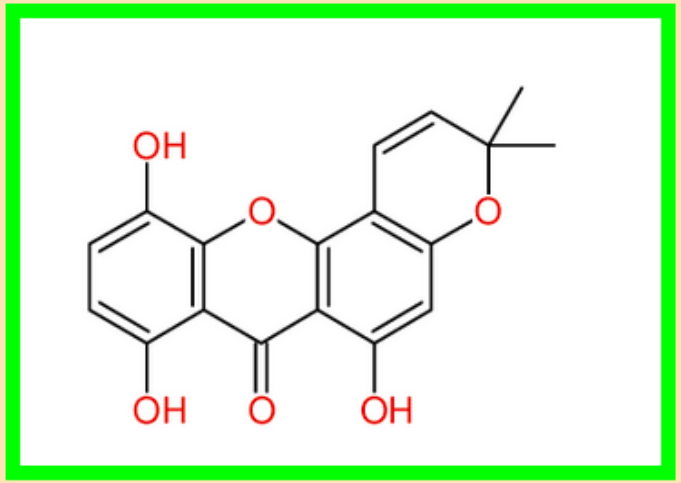
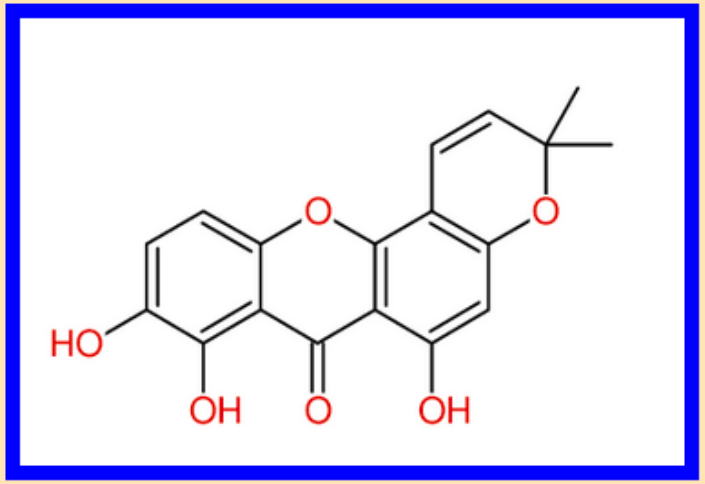
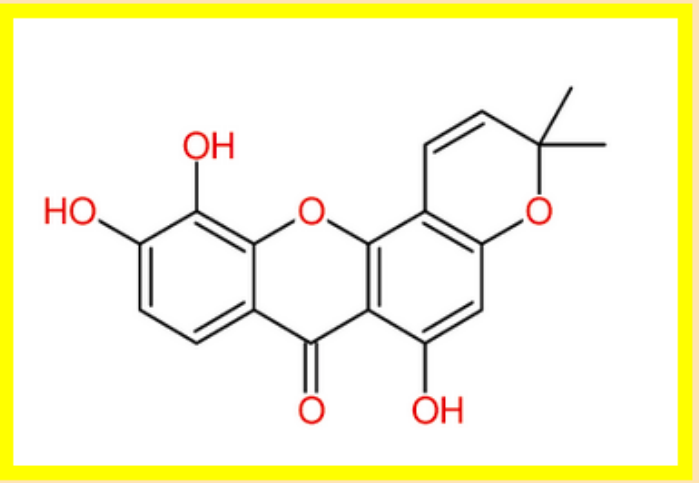
Eventually Symmetry Error: Same shiftvalue - Different environment



Strukturgenerator wird gestartet und erzeugt 4,348 alternative Strukturvorschläge
3 Strukturvorschläge sind bekannte Strukturen

Die Problematik mit diesem Datensatz wird erkannt, obwohl die identen Daten in der CSEARCH-Datenbank vorhanden sind !

Showing only structures either existing in CSEARCH or PUBCHEM

Proposal # 10	Proposal # 179	Proposal # 305
Similarity: 1.88ppm DNDDIKJWKDKBAB PUBCHEM	Similarity: 2.37ppm QFCIGJXDFQIUTI PUBCHEM	Similarity: 2.56ppm FSTNFJKGRSHPBO PUBCHEM
		

Strukturgenerator erzeugt 4,348 Alternativstrukturen – Ranking via Differenz „Experiment-Vorhersage“ – Abgleich mit CSEARCH und PUBCHEM, um real existierende Strukturen zu finden.

Scanning structural space:	**Completed**
CPU used:	**231.662 seconds**
Structures processed:	**4349**
Isomers / Non-Isomers:	**4034 / 315**
Structures with correct multiplicity:	**3281**
Structures available in CSEARCH or PUBCHEM:	**3**

CSEARCH-Database: Version 9.4.0 CSEARCH-Robot-Referee from 2017:06:10

Erkannte Probleme:

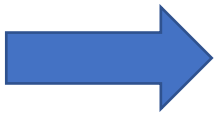
1 Position: Vorhersage → ca. 20 ppm Differenz; 108.6ppm(NN) bzw. 128.8ppm(HOSE)
Verschiebungswert bei 128.8 ppm kommt zweimal vor

Erklärungsmöglichkeiten:

- **1 bereits existierender Strukturvorschlag passt besser – Vergleich der publizierten Daten zeigt aber, dass doch signifikante Differenzen bestehen, Alternative auch nicht im Einklang mit einigen 2D-Messungen**

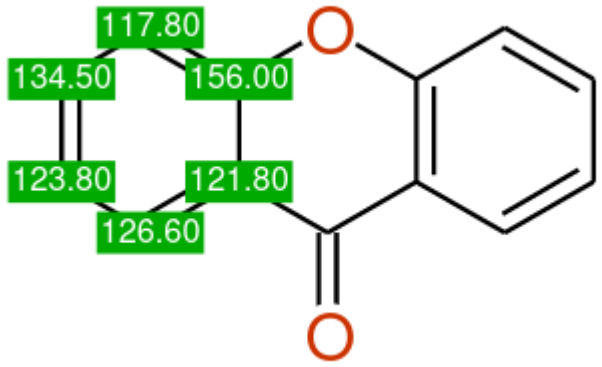
- **Wieso wurde das Signal bei 128.8ppm 2x zugeordnet ? Wieso ist bei diesem Signal 20ppm Abweichung zwischen den beiden Vorhersagen ? Tippfehler ?**

Überprüfung mittels Inkrementen (3. Vorhersagemethode neben HOSE+NN)



Inkrementberechnung

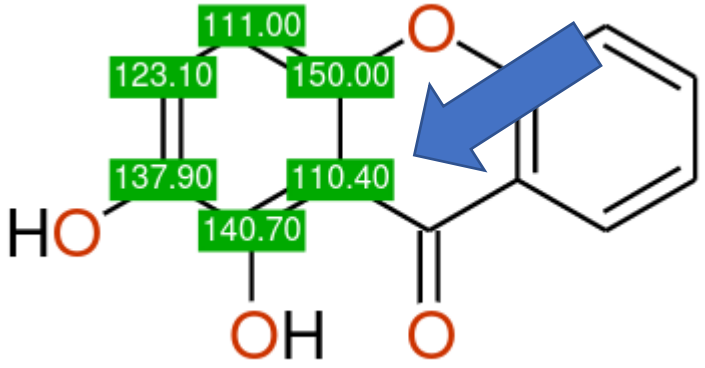
Grundkörper Xanthon - Literaturdaten

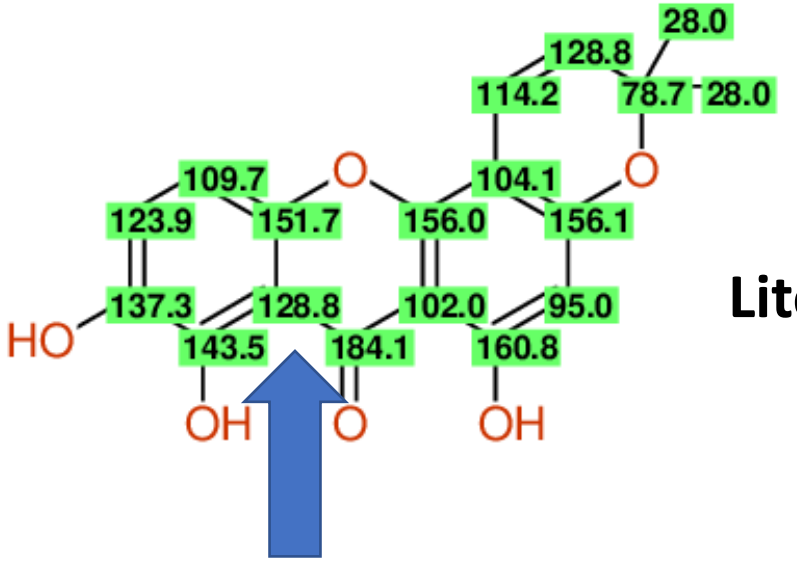


Vorhersage HOSE-Code: 128.8ppm wie Lit !
Vorhersage NN: 108.6ppm
Vorhersage Inkrement: 110.4ppm

Literatur: 128.8ppm – Signal jedoch 2x verwendet

2x Inkrement von -OH





Literaturdaten

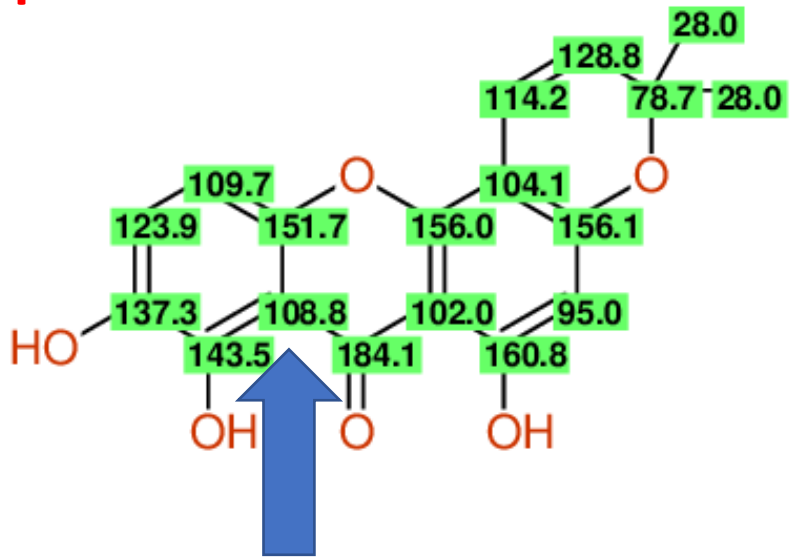
Zusammenfassung:

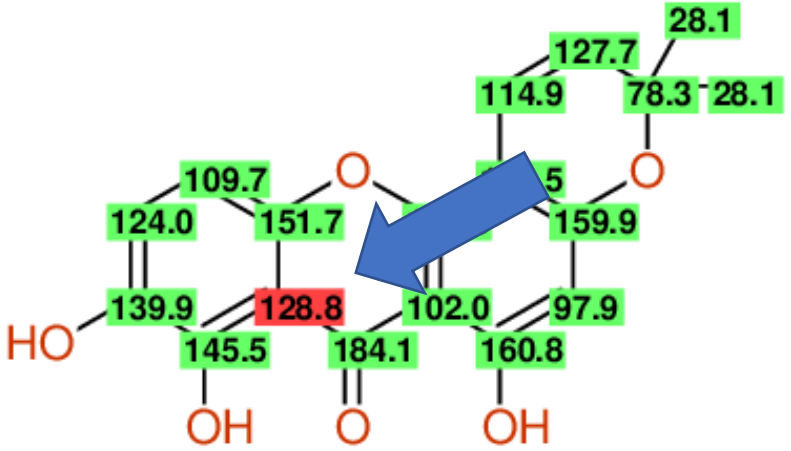
- **C13-Daten und Struktur – wie publiziert – inkompatibel; Struktur scheint korrekt**
- **Keine sinnvolle Alternativstruktur gefunden**
- **1 Signal bei 128.8ppm doppelt verwendet**
- **2 Vorhersagemethoden (NN + Inkremente) weisen auf Tippfehler beim quaternären C hin**
- **Leider keine Spektren in der „Supplementary Information“ verfügbar – die Fragestellung wäre mit wenig Aufwand zu entscheiden !**

Vermutlich 108.8ppm statt 128.8ppm → nochmals mit dem ‚korrigierten‘ Verschiebungswert bewerten lassen

Variante 2: Idente Struktur – 1 Verschiebungswert modifiziert (128.8 → 108.8 ppm)

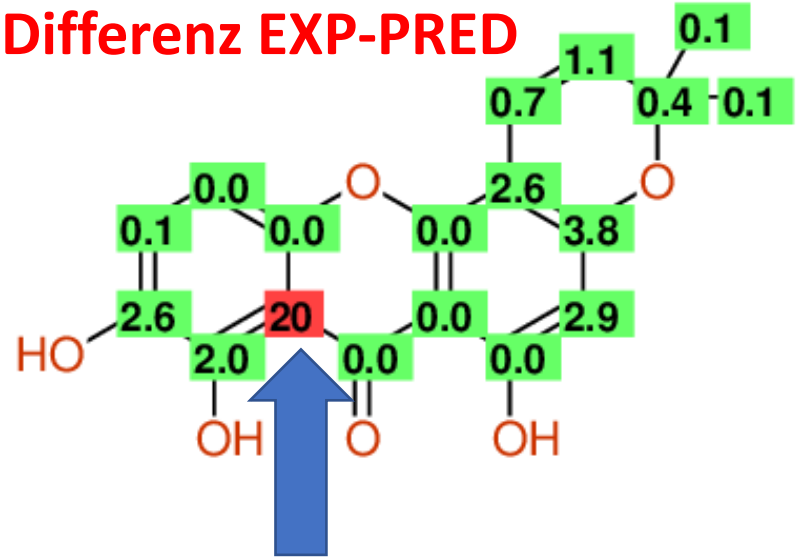
Exp-modifiziert





Idente Vohersage – nur andere Peakliste

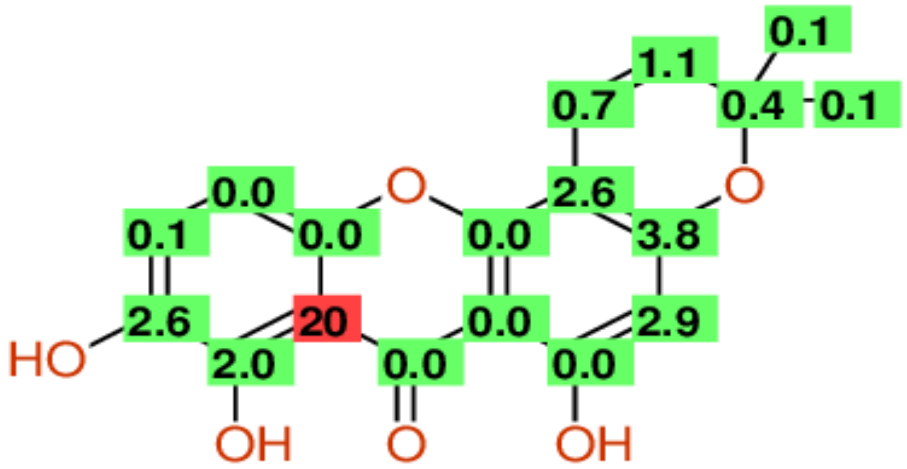
Differenz EXP-PRED



Carbon Number	Neural Network Prediction	HOSE-Code Prediction	Preferred Value from both Predictions	Experimental values	Difference (Exp-Pred/ppm)
1	161.6	160.8	160.8	160.8	0.0
2	99.8	97.9	97.9	95.0	2.9
3	161.9	159.9	159.9	156.1	3.8
4	99.9	101.5	101.5	104.1	2.6
5	153.5	156.0	156.0	156.0	0.0
6	148.3	151.7	151.7	151.7	0.0
7	108.9	109.7	109.7	109.7	0.0
8	126.4	124.0	124.0	123.9	0.1
9	143.8	139.9	139.9	137.3	2.6
10	148.4	145.5	145.5	143.5	2.0
11	108.6	128.8	128.8	108.8	20.0
12	184.9	184.1	184.1	184.1	0.0
13	100.4	102.0	102.0	102.0	0.0
14	77.2	78.3	78.3	78.7	0.4
15	128.2	127.7	127.7	128.8	1.1
16	120.3	114.9	114.9	114.2	0.7
17	27.8	28.1	28.1	28.0	0.1
18	27.8	28.1	28.1	28.0	0.1

Variante 2: Bewertung ergibt: Minor Revision - Begründung für diese Bewertung

Große Differenz zwischen HOSE-Code und NN-Vorhersage
1 Abweichung mit 20ppm zwischen Experiment und Vorhersage – Vorhersage ident zu vorher, weil idente Referenzdatenbank !



Strukturgenerator wird gestartet und erzeugt ebenfalls dieselben 4,348 Strukturvorschläge
3 Strukturvorschläge sind bekannte Strukturen – Bewertungsreihenfolge geringfügig anders, weil 1 Shiftwert modifiziert wurde (Bewertung: Differenz exp-pred !)

Zusammenfassung – Evaluierung von Globulixanthon C

Variante 1: Peakliste enthält 2x 128.8ppm:

Minor Revision, möglicher Symmetriefehler wird erkannt, Abweichung zwischen NN und HOSE-Vorhersage – Literaturdaten und Strukturvorschlag inkompatibel

Variante 2: Peakliste enthält 1x 128.8ppm, der zweite Wert wurde auf 108.8ppm modifiziert

Minor Revision, Abweichung zwischen NN und HOSE-Vorhersage, Abweichung zwischen Vorhersage und vorgegebenen Werten

Bewertung bleibt ident - die Inkonsistenzen werden in beiden Bewertungen korrekt und vollständig erkannt – obwohl die Referenzdaten (fragliche Daten bei beiden Bewertungen in der DB enthalten!) höchstwahrscheinlich falsch sind. Der Strukturgenerator erzeugt die identen 4,348 alternativen Strukturvorschläge – Reihenfolge unterschiedlich, weil experimentelle Peakliste geringfügig geändert wurde (128.8 → 108.8 ppm)

Nachdem „IDNMR“ laut Kollegen Schlörer (Sitzung IG „Kleine Moleküle“ vom 17.3.2021) sehr erfolgreich ist, schauen wir uns die idente Fragestellung mit dem „Quickcheck“ an:

Die Literaturdaten von „Globulixanthon C“ wurden ebenfalls korrekt aus der Literatur in NMRShiftDB übernommen, daher ist die Vergleichbarkeit gegeben

Quickcheck - Variante 1:

mit 2x128.8ppm – wie in der Literatur

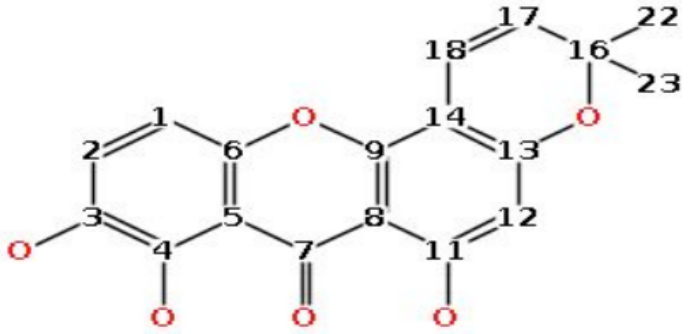
Resultat:

Quality report for carbon spectrum: nmrshiftdb2 quality check: 10 (accept)/reliability: excellent (Show full report)

Die höchstwahrscheinlich inkompatiblen Daten erhalten 10 von 10 Punkten, mit dem Qualitätsmerkmal „excellent“

Die Literaturdaten sind in NMRShiftDB – Konsistenzcheck erfolgreich, wird als Struktur- und Zuordnungsbeweis „verkauft“

nmrshiftdb2 Quality Report - ^{13}C



Atom	δ [ppm]	Deviation from prediction	Prediction		HOSE
			No. Spheres	No. shift values	Code
1	109.7	0.00	6.0	1.0	2D
2	123.9	0.00	6.0	1.0	2D
3	137.3	0.00	6.0	1.0	2D
4	143.5	0.00	6.0	1.0	2D
5	128.8	0.00	6.0	1.0	2D
6	151.7	0.00	6.0	1.0	3D
7	184.1	0.00	6.0	1.0	3D
8	102.0	0.00	6.0	1.0	3D
9	156.0	0.00	6.0	1.0	3D
11	160.8	0.00	6.0	1.0	3D
12	95.0	0.00	6.0	1.0	3D
13	156.1	0.00	6.0	1.0	3D
14	104.1	0.00	6.0	1.0	3D
16	78.7	0.30	6.0	3.0	3D
17	128.8	0.00	6.0	1.0	2D
18	114.2	0.00	6.0	1.0	2D
22	28.0	0.27	6.0	12.0	3D
23	28.0	Symmetric to atom 22	6.0	12.0	3D

Overall mark 10 (out of 1 to 10, 10 being best)
Mean deviation from prediction: 0.03 ppm → 0.02 Points
No. of red or missing shifts: 0.0 → 0.0 Points
No. of yellow shifts: 0.0 → 0.0 Points
Result: 10/accept

Quickcheck - Variante 2:

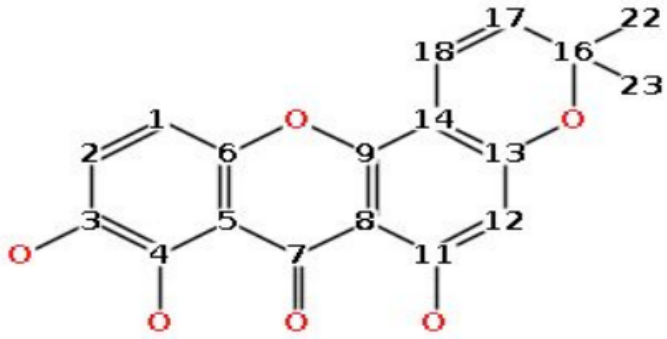
mit 1x128.8ppm
mit 1x108.8ppm

Resultat:

Quality report for carbon spectrum: nmrshiftdb2
quality check: 4 (revision)/reliability: excellent
(Show full report)

Die Variante, mit den höchstwahrscheinlich korrekten Daten erhält nur 4 von 10 Punkten, mit dem Qualitätsmerkmal „excellent“ (Es wird ja nur die Konsistenz und nicht die Richtigkeit bewertet !)

nmrshiftdb2 Quality Report - ^{13}C



Atom	δ [ppm]	Deviation from prediction	Prediction		HOSE
			No. Spheres	No. shift values	Code
1	109.7	0.00	6.0	1.0	2D
2	123.9	0.00	6.0	1.0	2D
3	137.3	0.00	6.0	1.0	2D
4	143.5	0.00	6.0	1.0	2D
5	108.8	20.00	6.0	1.0	2D
6	151.7	0.00	6.0	1.0	3D
7	184.1	0.00	6.0	1.0	3D
8	102.0	0.00	6.0	1.0	3D
9	156.0	0.00	6.0	1.0	3D
11	160.8	0.00	6.0	1.0	3D
12	95.0	0.00	6.0	1.0	3D
13	156.1	0.00	6.0	1.0	3D
14	104.1	0.00	6.0	1.0	3D
16	78.7	0.30	6.0	3.0	3D
17	128.8	0.00	6.0	1.0	2D
18	114.2	0.00	6.0	1.0	2D
22	28.0	0.27	6.0	12.0	3D
23	28.0	Symmetric to atom 22	6.0	12.0	3D

Overall mark 4 (out of 1 to 10, 10 being best)
Mean deviation from prediction: 1.21 ppm → 0.60 Points
No. of red or missing shifts: 1.0 → 5.0 Points
No. of yellow shifts: 0.0 → 0.0 Points
Result: 4/revision

Quickcheck/NMRShiftDB - Zusammenfassung:

Die besser passenden Daten erhalten 4 von 10 Punkten, die schlechter passenden Daten 10 von 10 Punkten

Warum?

- **Nur 1 Vorhersagemethode – nur HOSE-Code**
- **Die verwendete Vorhersagemethode reproduziert exakt den Datenbankinhalt inklusive potentieller Fehler aufgrund des dahinter stehenden mathematischen Modells**
- **Es liegt hier keine Qualitätskontrolle im eigentlichen Sinn vor, bloß ein Konsistenzcheck – der vermutlich falsche Datensatz zeigt die bessere Konsistenz, da dieser Datensatz im Referenzmaterial ist.**
- **2 Spalten (experimentelle Shiftwerte und Abweichungen) in einer Tabelle, welche nach 5 Jahren immer noch „currently experimental“ ist, zeigt eine gewisse „Schlichtheit“ im Verständnis von „korrekt“ und „konsistent“.**
- **Keine grafische Aufbereitung der Daten**

Zielsetzung für Projekt „ Computer-Assisted Peer-Reviewing“

1) **Korrekte und validierte Datensätze mit hoher Qualität als zukünftiges, vertrauenswürdiges Referenzmaterial – zB. als Wissensbasis für weitere Entwicklungen im Bereich ‚Vorhersagemodelle‘**

2) **Vollständige Strukturcharakterisierung und somit Vermeidung von falschen Strukturen**

Zielsetzung für Projekt „ Computer-Assisted Peer-Reviewing“

1) Korrekte und validierte Datensätze mit hoher Qualität als zukünftiges, vertrauenswürdiges Referenzmaterial – zB. als Wissensbasis für weitere Entwicklungen im Bereich ‚Vorhersagemodelle‘

2) Vollständige Strukturcharakterisierung und somit Vermeidung von falschen Strukturen

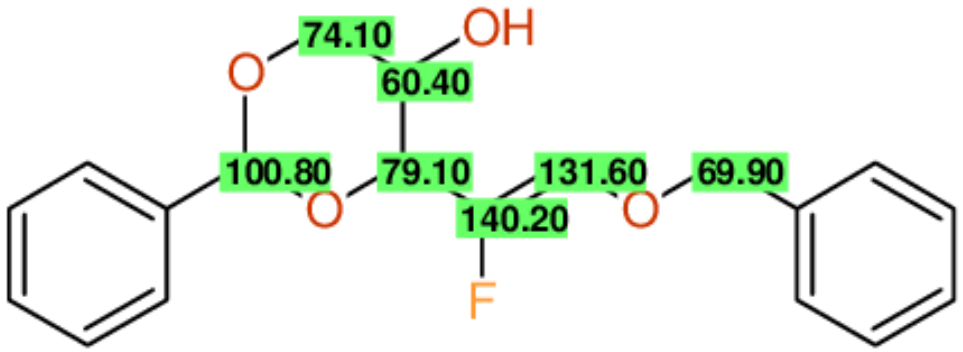
Sitzung der IG „Kleine Moleküle“ vom 17. März 2021

- Aufruf zur Mitarbeit in einem Arbeitskreis zur Definition von Standards zum Publizieren von NMR-Daten

- Herantreten an die Verlage um diesen Standard in Publikationen umzusetzen

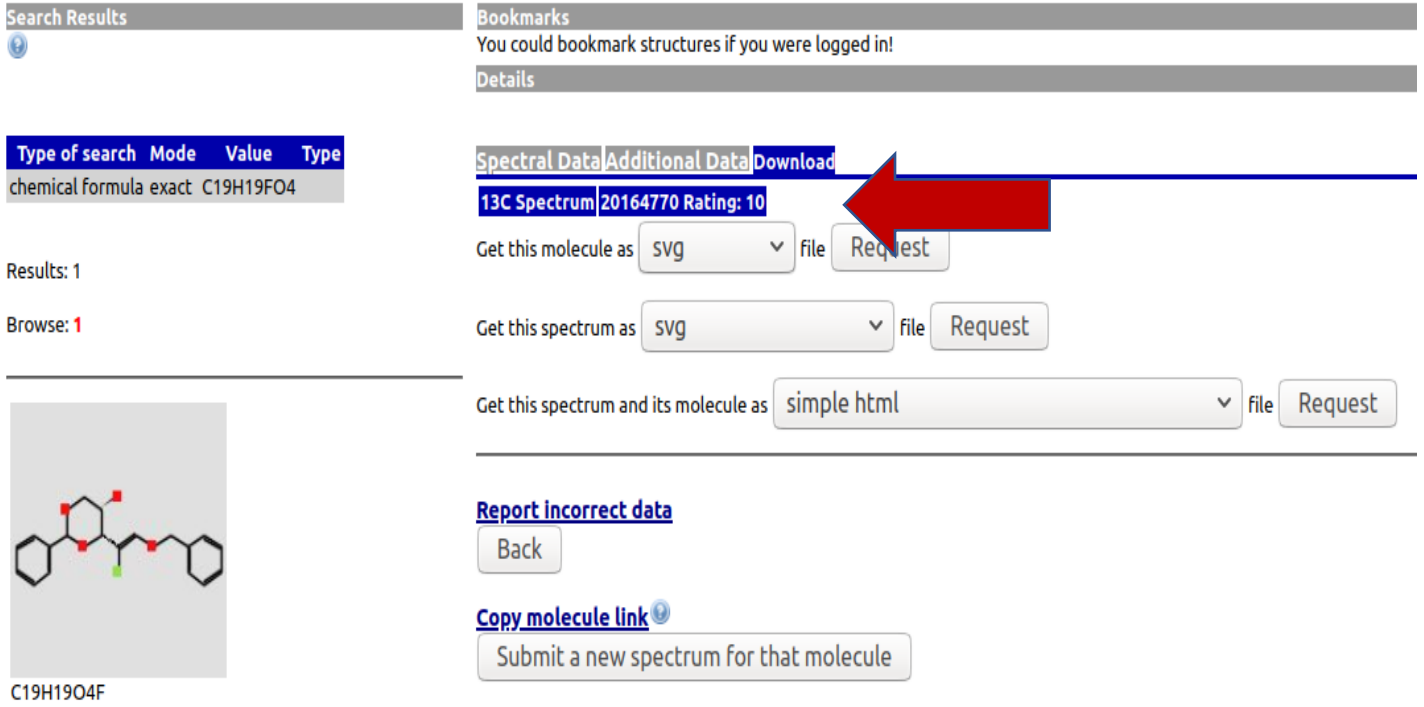
CSEARCH-Robot-Referee

Reject – weil Linien fehlen



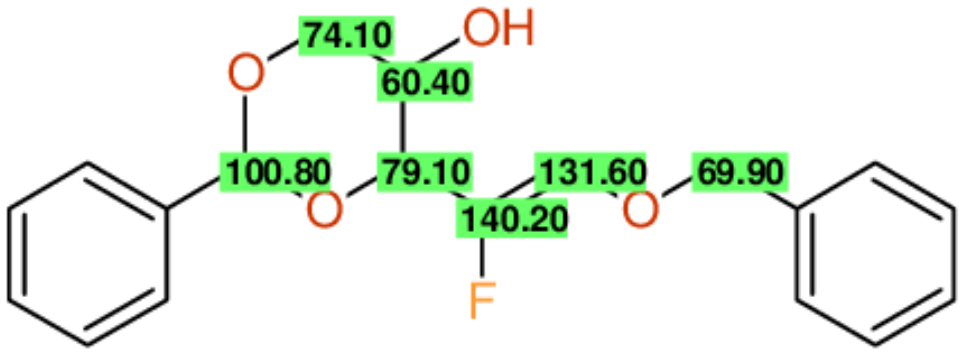
Quickcheck/NMRShiftDB:

10 von 10 Punkten



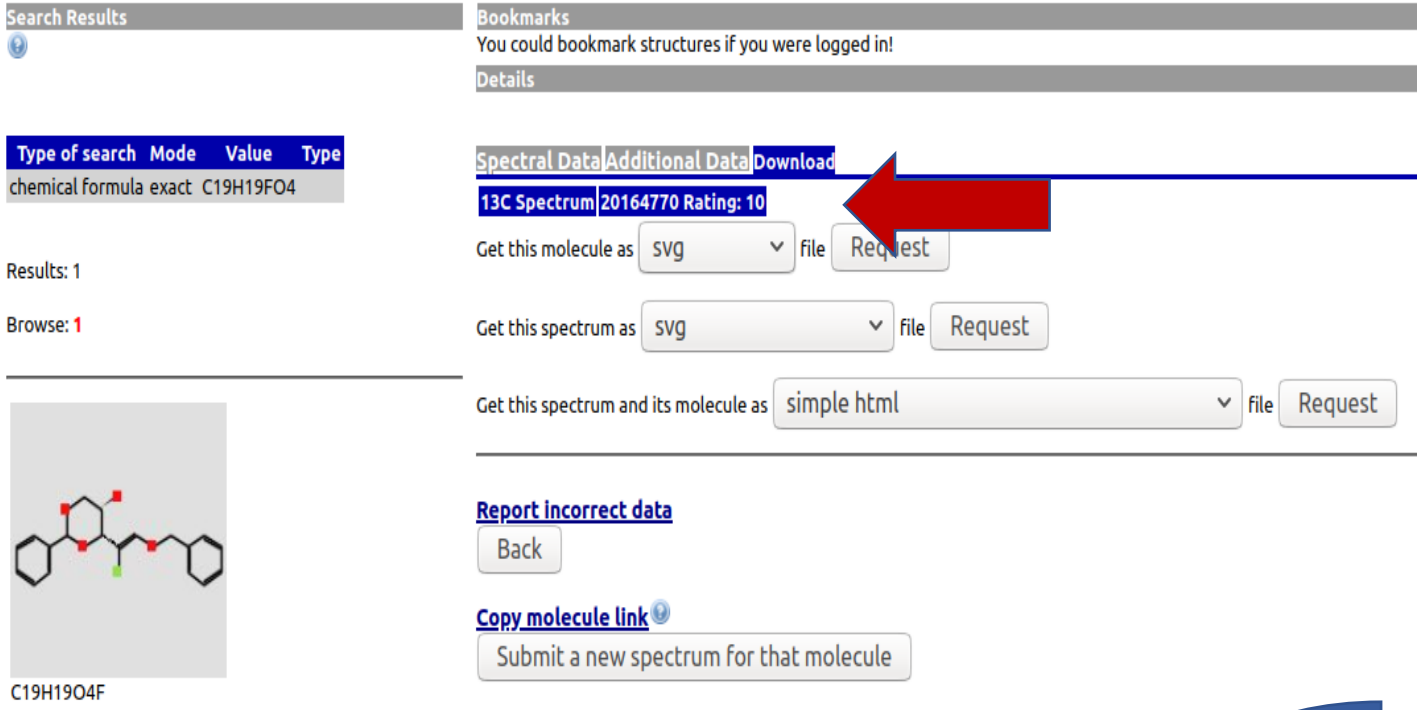
CSEARCH-Robot-Referee

Reject – weil Linien fehlen



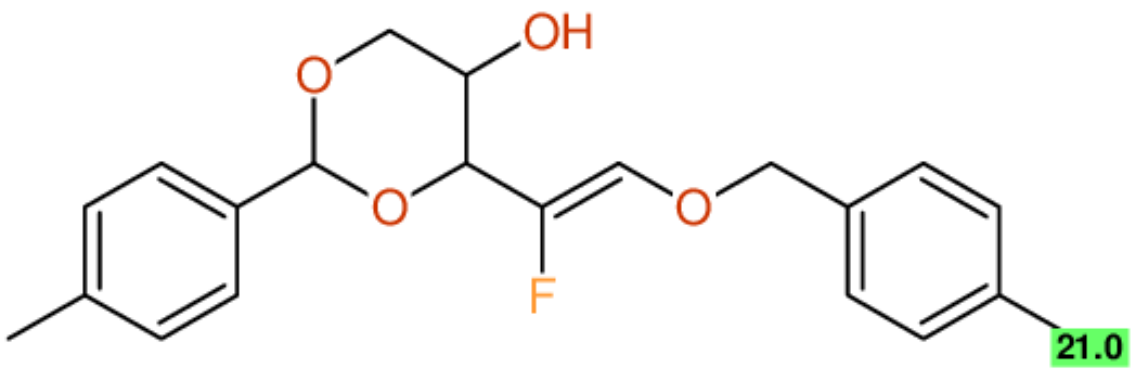
Quickcheck/NMRShiftDB:

10 von 10 Punkten



Sie werden‘s nicht glauben, aber Strukturbeweise mittels NMRShiftDB lassen sich noch weiter „optimieren“ !

Quickcheck gives:



Atom	δ [ppm]	Deviation from prediction	Prediction		HOSE
			No. Spheres	No. shift values	Code
		-	-	-	-
		-	-	-	-
		-	-	-	-
		-	-	-	-
		-	-	-	-
		-	-	-	-
		-	-	-	-
		-	-	-	-
10	-	-	-	-	-
11	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	-	-	-	-	-
	21.0	0.00	6.0	1.0	2D
	-	-	-	-	-

Overall mark 10 (out of 1 to 10, 10 being best)

Mean deviation from prediction: 0.00 ppm → 0.00 Points

No. of red or missing shifts: 20.0 → 100.0 Points

Resultat: 10 von 10 Punkten = Accept

Reliability: Excellent

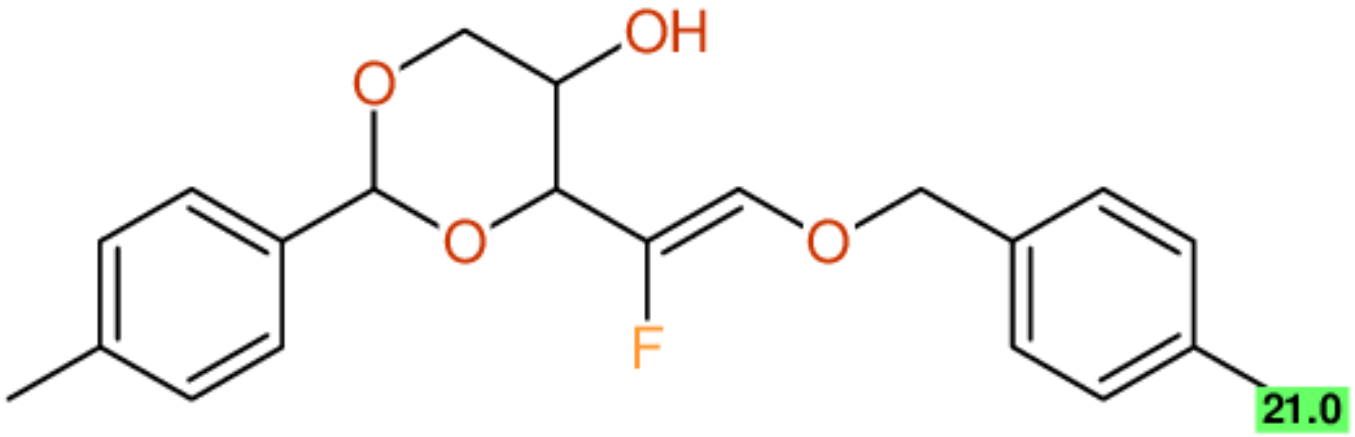
No. of yellow shifts: 0.0 → 0.0 Points

Result: 10/accept

Mehr als 10 Punkte von 10 Punkten geht ja nicht

Eine bessere „Reliability“ als „excellent“ gibt’s ja auch nicht mehr !

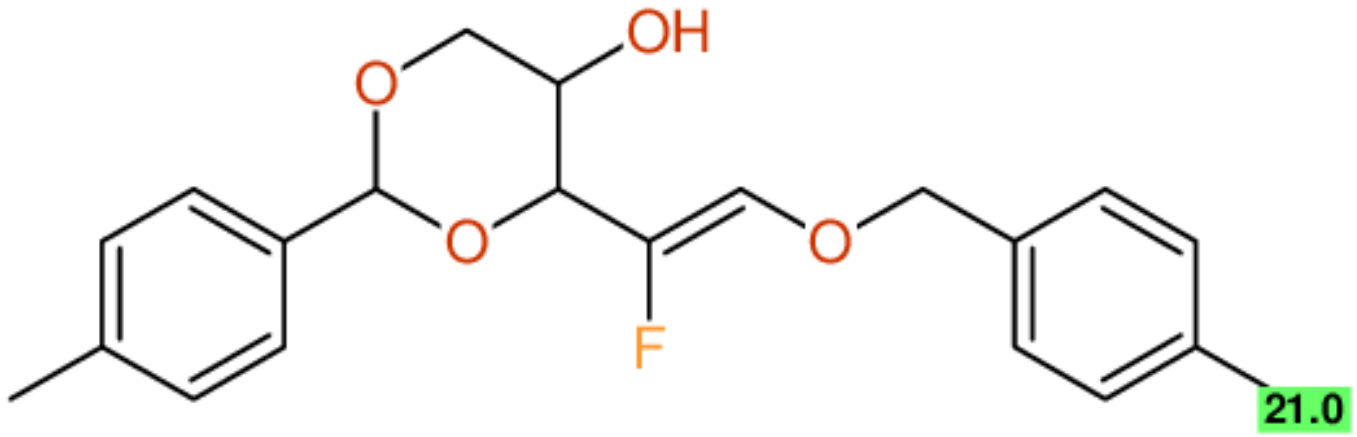
Eine Linie bei 21ppm ist also ein eindeutiger Strukturbeweis laut „Quickcheck“ für untenstehende Struktur



Mehr als 10 Punkte von 10 Punkten geht ja nicht

Eine bessere „Reliability“ als „excellent“ gibt’s ja auch nicht mehr !

Eine Linie bei 21ppm ist also ein eindeutiger Strukturbeweis laut „Quickcheck“ für untenstehende Struktur



Sorry, aber das ist zum

Ich danke Ihnen für Ihre Aufmerksamkeit !

Die Powerpoint-Präsentation steht demnächst unter
http://c13nmr.at
zum Download zur Verfügung

Weitere Beispiele unter http://c13nmr.at/literature