Substitution Patterns in Aromatic Rings by Increment Analysis (SPARIA) – Model Development and Application to Natural Organic Matter

E. M. Perdue[*], N. Hertkorn[a], A. Kettrup[a]

[*] *School of Earth and Atmospheric Sciences,*
Georgia Institute of Technology,
Atlanta, Georgia, 30332, USA
mperdue@eas.gatech.edu; Phone: 1-404-894-3942; FAX: 1-404-894-5638

[a] *GSF-Forschungszentrum für Umwelt und Gesundheit*
Institut für Ökologische Chemie,
85758 Neuherberg, Germany
hertkorn@gsf.de; Phone: +4989-3187-2834; FAX: +4989-3187-2705

ABSTRACT

The tables and figures in this document are provided as supporting information for the manuscript whose title is given above. Each table or figure is cross-referenced in the manuscript.

Using Incremental Chemical Shifts of Substituents on Benzene Rings

As a simple illustration of the use of the data in Table 1 for prediction of chemical shifts of ^{1}H and ^{13}C in aromatic compounds, consider 3,5-dimethoxy-4-hydroxybenzoic acid, which contains two equivalent aromatic C—H bonds. The observed chemical shifts of ^{1}H and ^{13}C are 7.19 ppm and 106.9 ppm, respectively. The calculated chemical shifts of ^{1}H and ^{13}C, using data from Table 1, are shown in Table S-1:

Table S-1. Forward prediction of the chemical shifts of ^{1}H and ^{13}C in 3,5-dimethoxy-4-hydroxybenzoic acid using increment analysis.

	Reference	o–COOH	m-H	p–OCH_3	m'–OH	o'–OCH_3	Prediction
^{1}H	7.26	+0.85	0.00	-0.44	-0.12	-0.48	7.07
^{13}C	128.5	+1.6	0.0	-8.1	+1.6	-15.0	108.6

The Algorithm Used in the Forward Mode of SPARIA

A simple computer program (in Pascal) for generating the database of 32768 substitution patterns and their chemical shifts of ^{1}H and ^{13}C is given in Table S-2. The declarations of

Table S-2. A Pascal program for generating the database of substitution patterns and chemical shifts that are used in SPARIA.

```
PROGRAM Patterns;

CONST
  Groups = 8;     {The number of substituents used in SPARIA}
  Positions = 5;  {The number of substituted ring positions}
  Total = 32768;  {The maximum number of permutations is 8^5}

  {List of Substituents}
  Name: ARRAY [0..Groups-1] OF STRING[6] =
  ('H', 'C2H5', 'CH=CH2', 'COOH', 'CO2CH3', 'COC2H5', 'OCH3',
'OH');

  {The starting 1H and 13C chemical shifts for benzene}
  Start_H = 7.26;
  Start_C = 128.5;

  {The 1H chemical shift factors for the substituents}
  Ortho_H: ARRAY [0..Groups-1] OF REAL =
    ( 0.00, -0.15,  0.06,  0.85,  0.71,  0.63, -0.48, -0.56 );
  Meta_H:  ARRAY [0..Groups-1] OF REAL =
    ( 0.00, -0.06, -0.03,  0.18,  0.11,  0.13, -0.09, -0.12 );
  Para_H:  ARRAY [0..Groups-1] OF REAL =
    ( 0.00, -0.18, -0.10,  0.25,  0.21,  0.20, -0.44, -0.45 );

  {The 13C chemical shift factors for the substituents}
  Ortho_C: ARRAY [0..Groups-1] OF REAL =
    ( 0.00,  -0.6,  -1.8,   1.6,   1.0,   0.2, -15.0, -12.6 );
  Meta_C:  ARRAY [0..Groups-1] OF REAL =
    ( 0.00,  -0.1,  -1.8,  -0.1,   0.0,   0.2,   0.9,   1.6 );
  Para_C:  ARRAY [0..Groups-1] OF REAL =
    ( 0.00,  -2.8,  -3.5,   4.8,   4.5,   4.2,  -8.1,  -7.6 );

VAR
  I, J, K: BYTE;
  L, M, N: WORD;
  Shift_H, Shift_C: REAL;
  DevO: TEXT;

BEGIN
  ASSIGN (DevO, 'PATTERNS.OUT');
  REWRITE (DevO);
  FOR L:=0 TO Total-1 DO
  BEGIN
    M:=L;
    Shift_H:=Start_H;
    Shift_C:=Start_C;
    FOR I:=1 TO Positions DO
    BEGIN
      N:=1;
      FOR J:=1 TO (Positions-I) DO N:=N*8;
      K:=M DIV N;
      CASE I OF
        1,5: BEGIN
               Shift_H:=Shift_H+Ortho_H[K];
               Shift_C:=Shift_C+Ortho_C[K];
             END;
        2,4: BEGIN
               Shift_H:=Shift_H+Meta_H[K];
               Shift_C:=Shift_C+Meta_C[K];
             END;
        3  : BEGIN
               Shift_H:=Shift_H+Para_H[K];
               Shift_C:=Shift_C+Para_C[K];
             END;
      END;
      WRITE(DevO,Name[K]:10);
      M:=M-K*N;
    END;
    WRITELN(DevO,Shift_H:10:2,Shift_C:10:1);
  END;
  CLOSE (DevO);
END.
```

constants and variables are straightforward. In the main program block, each of the 32,768 (8^5) possible substitution patterns is assigned an index in the range of 0 to 32,767 in the outer "L"

loop. Each decimal-based index is then expressed as a five digit, base-eight (octal) number. More generally, the number of digits in the number must equal the number of ring positions (5) and the base must equal the number of substituents (8). Each digit of the resulting five-digit octal number corresponds to an element of the Name array at one of the five ring positions. The individual digits of the octal number are extracted sequentially in the "I" loop and are the used to index the Name array and the arrays for incremental chemical shifts. Consider, for example, a substitution pattern whose decimal index is 12345. When expressed as an octal number, 12345 becomes:

$$12345 \rightarrow 3x8^4 + 0x8^3 + 0x8^2 + 7x8^1 + 1x8^0 = 30071_8$$

The substituents can then be assigned to the ring positions as follows:

Position	Ortho	Meta	Para	Meta'	Ortho'
Digit	3	0	0	7	1
Substituent	COOH	H	H	OH	C2H5

Once a substituent is assigned to a ring position, its contributions to the chemical shifts of ^{1}H and ^{13}C are added to the initial values for benzene. Each substitution pattern and set of predicted chemical shifts are written to an output file. The program executes in 0.3 seconds on a 3 GHz IBM personal computer.

Compounds Used to Test Forward and Inverse Predictions of the SPARIA Model

The 29 compounds used to test the forward and inverse modes of SPARIA are listed in Table S-3 and their structures are given in Figure 4. Collectively, these compounds contain most of the substituents that are used in SPARIA, and they are often invoked as likely structural subunits in natural organic matter.

Table S-3**.** Compounds used to test forward and inverse predictions of the SPARIA Model.

1	benzoic acid	**11**	4-methoxyacetophenone	**21**	4-hydroxy-3-methoxyacetophenone
2	R-mandelic acid	**12**	o-coumaric acid	**22**	homovanillic acid
3	alpha-methylcinnamic acid	**13**	m-coumaric acid	**23**	ferulic acid
4	resorcinol	**14**	p-coumaric acid	**24**	3,4-dimethoxyacetophenone
5	2-methoxyphenol	**15**	coumarin-3-carbonic acid	**25**	gallic acid
6	o-hydroxybenzoic acid	**16**	1,2,4-trimellitic acid	**26**	3,5-dimethoxy-4-hydroxyacetophenone
7	m-hydroxybenzoic acid	**17**	protocatechoic acid	**27**	syringic acid
8	p-hydroxybenzoic acid	**18**	2,4-dihydroxybenzaldehyde	**28**	3,4,5-trimethoxyacetophenone
9	phthalic acid	**19**	caffeic acid	**29**	pyromellitic acid
10	terephthalic acid	**20**	vanillin		

In Figure S-1, the 100 experimental NMR peaks for the 29 compounds in Table S-3 and Figure 4 are superimposed on the predicted peaks for the 16,640 unique substitution patterns.

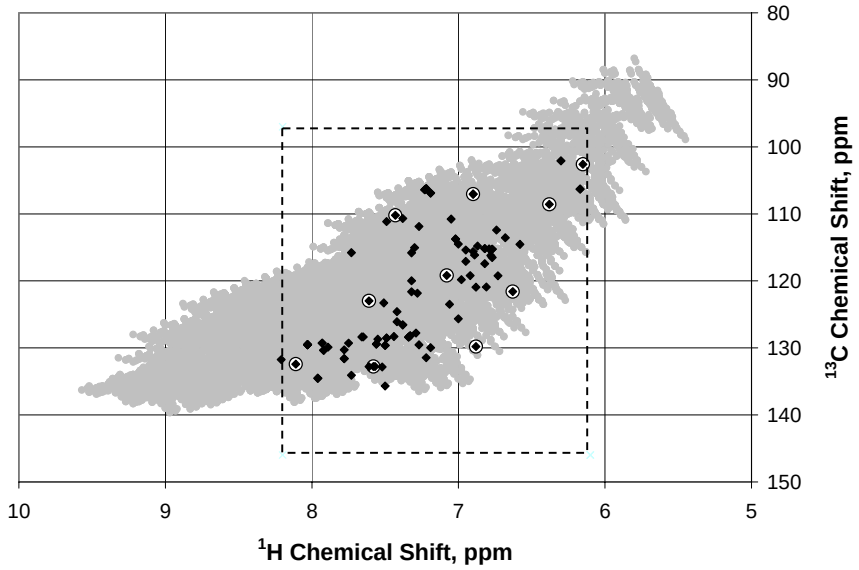


Figure S-1. The NMR peaks of 29 compounds used to evaluate the forward and inverse modes of SPARIA (see Figure 4 and Table S-3).

The highlighted subset of 10 peaks in Figure S-1 is used to optimize the inverse mode of SPARIA – specifically the size of the target window of chemical shift for ^{1}H and ^{13}C. It is noteworthy that all the peaks fall within the rectangular area within which most peaks for NOM and related materials are found.

More than 50% of the compounds in Table S-3 and Figure 4 were necessarily modelled using surrogate structures containing the eight substituents used in the forward mode of SPARIA. A few examples are shown in Figure S-2. In some instances, the surrogate structures are quite similar to the actual structures that are being modeled; however, compounds containing complex side chains were not well represented by surrogate structures.

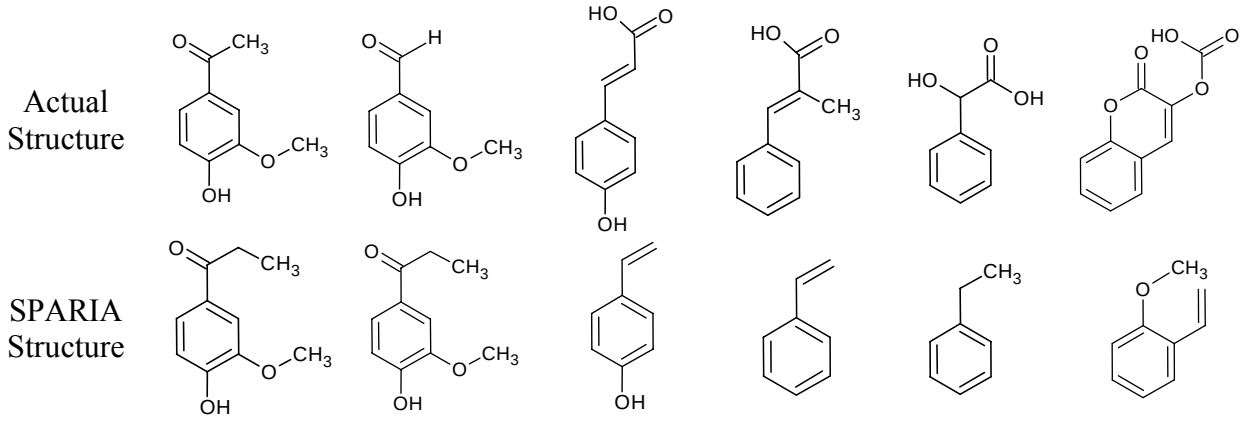


Figure S-2. Actual and surrogate structures used to test the accuracy of the forward mode in SPARIA.

An Initial Effort to Refine the Forward Mode of SPARIA

In the course of using SPARIA in the forward mode, it became clear that predictions of the chemical shift of ^{1}H were often poor in molecules containing COOH groups that were ortho to each other (see **9**, **16** and **29** in Figure 4). The incremental chemical shifts of COOH were optimized for such groups using the solver tool in Microsoft Excel.

Table S-4. Incremental chemical shifts for selected substituents on aromatic rings – ^{1}H and ^{13}C incremental chemical shifts are relative to 7.26 ppm and 128.5 ppm, respectively.

	^{1}H Incremental Chemical Shift			^{13}C Incremental Chemical Shift		
	ortho	**meta**	**para**	**ortho**	**meta**	**para**
—COOX *	0.46	-0.13	0.34	2.4	-0.2	3.8
—CH=CHX *	0.39	0.08	0.13	-0.4	-0.5	0.9

* COOX represents o-dicarboxylic acids and CH=CHX represents –CH=CH—COOH in limited tests to improve the forward predictions of SPARIA.

Similarly, it was observed that predictions of the chemical shift of ^{13}C were poor in structures containing the –CH=CH—COOH group (see **3**, **12**, **13**, **14**, **19**, and **23** in Figure 4). This resonance electron-withdrawing substituent is not adequately represented by the resonance electron-donating $–CH=CH_2$ substituent, which is used in SPARIA. A new set of incremental chemical shifts was obtained using the solver in Microsoft Excel for such structures. The optimized incremental chemical shifts for torsionally strained –COOH and for –CH=CH—COOH are given in Table S-4.

When SPARIA uses the modified incremental chemical shifts from Table S-4, the agreement between observed and calculated chemical shifts improves considerably (see Table S-5). When compared with the RMSE values for the standard SPARIA model, the RMSE of predicted ^{1}H chemical shifts is improved dramatically by use of the optimized parameters for –COOX; however, this adjustment had little effect on the RMSE of predicted ^{13}C chemical shifts. Torsional strain from steric interaction between ortho-COOH groups weakens mesomeric effects but has less impact on polar electron withdrawal. Furthermore, a variable alignment of the carbonyl (C=O) bond vector alters the chemical shift anisotropy experienced by neighboring atoms, and both these interactions more strongly affect aromatic protons than carbon atoms. Conversely, the RMSE of predicted ^{13}C chemical shifts is improved dramatically by use of the optimized parameters for –CH=CH–COOX; however, this adjustment had little effect on the RMSE of predicted ^{1}H chemical shifts. When both sets of corrected chemical shifts are used, the RMSE for predicted chemical shifts of ^{1}H is actually slightly better than the RMSE for predictions generated by the ACD/HNMR Predictor 5.0. Even with these refinements, the forward mode of SPARIA still could not match the predictions of the ACD/CNMR Predictor 5.0 for ^{13}C chemical shifts.

Given that the ultimate goal is to apply SPARIA to natural organic matter, for which it is impossible to know if/when standard incremental chemical shifts of substituents should be

115 modified, the forward mode of SPARIA was implemented in this paper using the standard
116 incremental chemical shifts of substituents that are in Table 1.
117

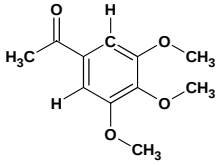
Table S-5. Linear regression analysis and root mean square error (in ppm) for chemical shifts of ^{1}H and ^{13}C in the 29 compounds in Figure 4 and in Table S-3.

	Regression Parameters for ^{1}H				**Regression Parameters for ^{13}C**			
Model *	RMSE	Intercept	Slope	R^2	RMSE	Intercept	Slope	R^2
ACD	0.23	0.69	0.92	0.79	1.52	2.32	0.99	0.98
SPARIA - Standard	0.35	-1.55	1.21	0.73	2.27	13.27	0.89	0.94
SPARIA - COOX	0.25	-0.31	1.03	0.81	2.22	13.91	0.89	0.94
SPARIA - CH=CHX	0.33	-1.41	1.19	0.75	1.94	12.77	0.90	0.96
SPARIA – Both X	0.22	-0.17	1.02	0.83	1.91	12.45	0.90	0.96

* COOX represents o-dicarboxylic acids and CH=CHX represents –CH=CH—COOH in limited tests to improve the forward predictions of SPARIA. These modifications were conducted separately and together (Both X).

118
119 **Analysis of all 80 "Hits" that were obtained by the Inverse Mode of SPARIA for 3,4,5-**
120 **trimethoxyacetophenone**
121
122 Table 3 contains an illustrative analysis of 11 of the 80 matching substitution patterns that were
123 obtained 3,4,5-trimethoxyacetophenone using the inverse mode of SPARIA. If all eighty "hits"
124 are averaged, the results in Table S-6 are obtained.
125

Table S-6. Inverse mode of SPARIA using all 80 matching substitution patterns for an aromatic C—H group in 3,4,5-trimethoxyacetophenone (**28** in Figure 4).

	Observed Peak, ppm		**Peak Window, ppm**		**Structure (for reference only)**
Input to SPARIA	**$\delta^{1}H$**	**$\delta^{13}C$**	**$\Delta(\delta^{1}H)$**	**$\Delta(\delta^{13}C)$**	
	7.23	106.40	0.10	1.00	

Classes of Substituents (Probability)		**Ortho**	**Meta**	**Para**	**Meta'**	**Ortho'**
	COR	1.000	0.625	0.000	0.625	0.000
	R	0.000	0.362	0.038	0.362	0.000
	OR	0.000	0.013	0.962	0.013	1.000

Error (SPARIA – Actual)		**Ortho**	**Meta**	**Para**	**Meta'**	**Ortho'**
	COR	0.000	0.625	0.000	0.625	0.000
RMSE = 0.39	**R**	0.000	-0.638	0.038	0.362	0.000
	OR	0.000	0.013	-0.038	-0.987	0.000

126
127 The results in Table S-6 confirm that the 11 substitution patterns used in Table 3 are
128 representative of the entire set of "hits" that were obtained for 3,4,5-trimethoxyacetophenone.
129

Optimizing the Width of the Target Window for the Inverse Mode of SPARIA

The variation of the *average* RMSE for predicted substitution patterns with the half-width of the ^{13}C window for the ten peaks that are highlighted in Figure S-2 is given in Figure 6. Figure S-3 contains this result (dashed line) and the ten individual RMSE curves. In all calculations, the half-width of the ^{1}H window is 0.1 times that of the ^{13}C window. Some RMSE's never pass through a minimum, or they pass through a minimum either at smaller or greater window size. The standard half-width of 1 ppm for ^{13}C and 0.1 ppm for ^{1}H is clearly a compromise choice.

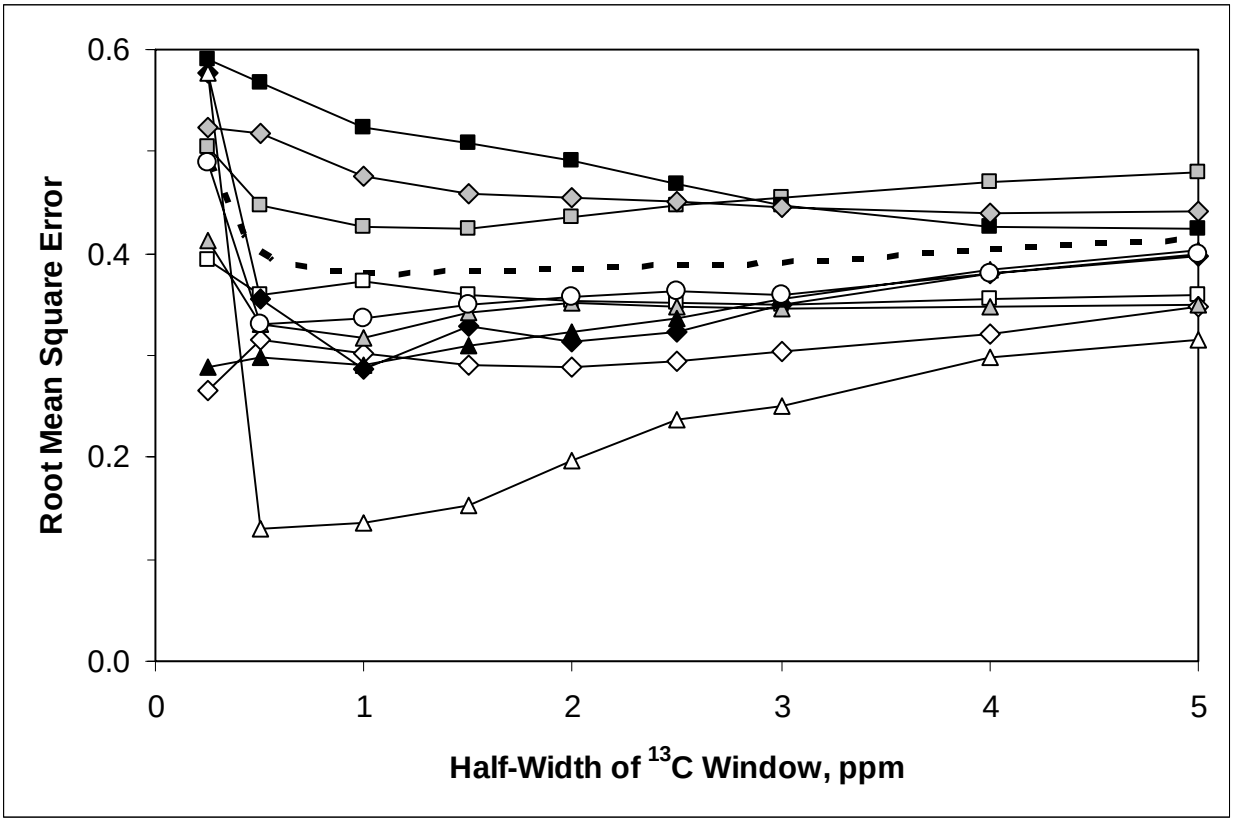


Figure S-3. Error analysis of predicted substitution patterns versus half-width of the chemical shift window of ^{13}C ($\Delta(\delta\ ^{13}C)$).

Comprehensive Error Analysis of the Inverse Mode of SPARIA

For each predicted substitution pattern, the error of the prediction is calculated at each of the five ring positions for each of the three classes of substituents, i.e., a total of fifteen individual errors, and the overall error in the prediction is expressed as the root mean square of the 15 individual errors. The frequency and cumulative distributions of errors from the inverse mode of SPARIA for the 100 peaks used previously in this paper to test the forward mode of SPARIA are given in Figure S-4. Unlike the corresponding results in Figure 7, in which only cumulative frequency distributions are provided and for which errors for ortho and ortho' positions were consolidated, as were errors for meta and meta' positions, Figure S-4 gives frequency distributions and cumulative frequency distributions for all three classes of substituents at all five ring positions.

151
152 The frequency plots on the left side of Figure S-4 reveal that the error distribution for predicted
153 classes of substituents at ortho, ortho', and para positions is unimodal and centered on zero. In
154

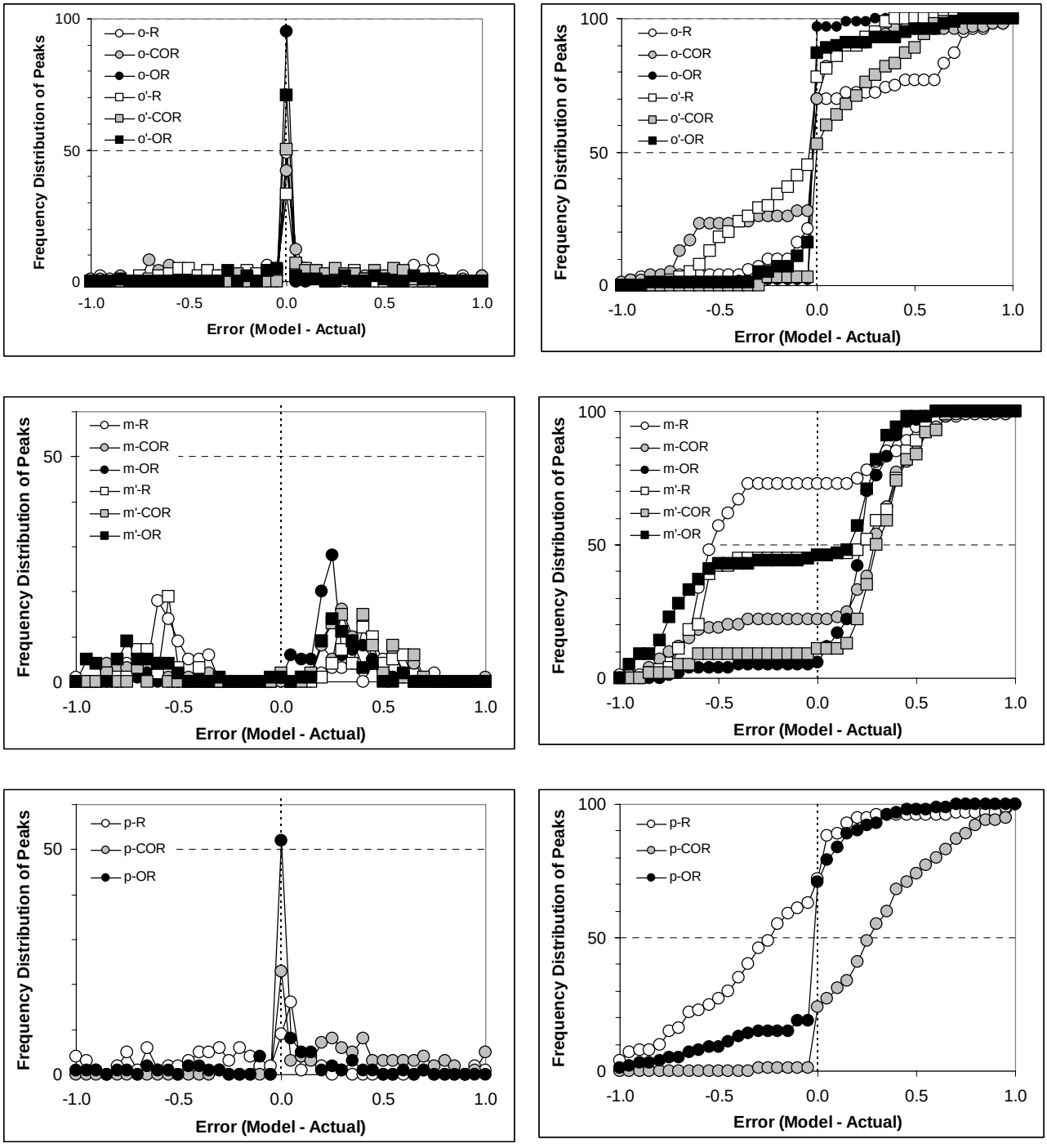


Figure S-4. Errors in predictions of the inverse mode of SPARIA, including frequency distributions and cumulative distributions of error.

155

contrast, a bimodal distribution of errors is obtained for predicted classes of substituents at meta and meta' positions. As discussed in the main paper, the maxima of the bimodal distribution are reasonably close to the most probable positive and negative errors for totally random predictions.

Some additional details are visible in the cumulative error plots on the right side of Figure S-4. Because of the highly accurate predictions of –OR substituents in the ortho and ortho' positions, compensatory errors are found in the predictions of –R and –COR groups in the ortho and ortho' positions. The moderate underestimation of –COR groups in the ortho position, for example, is almost exactly balanced by a corresponding overestimation of –R groups in that position. An opposite, but also compensatory, pattern of errors is found in the predictions of –R and –COR groups in the ortho' position.

Another view of the predictive capabilities of the inverse mode of SPARIA is obtained by rounding the probability of occurrence of a class of substituents to zero or one. The most probable class of substituents is assigned a probability of one, and the other two classes of substituents are assigned probabilities of zero. The rationale for this approach is that a class of substituents must ultimately be present or absent in a substitution pattern. Errors are again calculated as (SPARIA – Actual), so errors must now equal -1, 0, or 1. The resulting probability distribution of errors is given in Figure S-5.

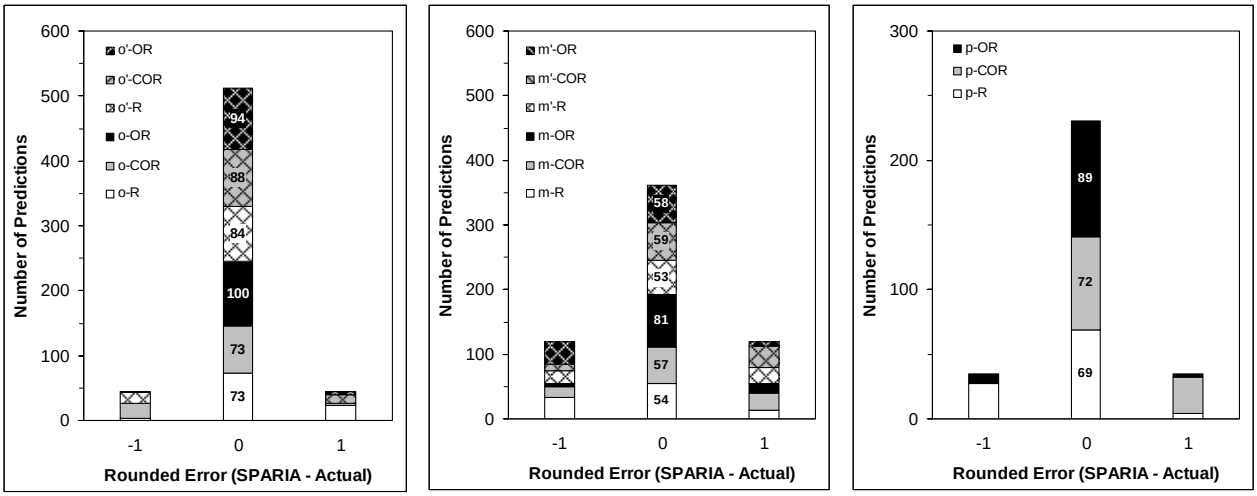


Figure S-5. Overall summary of the performance of the inverse mode of SPARIA for three classes of substituents at ortho, meta, and para ring positions, when errors are rounded (see text for discussion). Numbers inside the data bars represent the number of correct predictions for a particular class of substituent and ring position (maximum = 100).

The overall distributions of error are now necessarily symmetrical; however, the error distributions for individual classes of substituents remain asymmetrical. The overall percentages of correct predictions (using rounded probabilities) from the inverse mode of SPARIA are 85%, 60%, and 77%, respectively, at the ortho, meta, and para ring positions. When all such results are combined, the global percentage of correct predictions is 74%.