Supporting Information

Table 1: Overlapping versus contradictory regioselectivity signal
Diagonal: Number of substrates in each set.
Upper-triangle: Number of substrates in both sets with at least one overlapping observed SOM.
Lower-triangle: Number of substrates in both sets with no overlapping observed SOMs.

Isozyme	1A2	2A6	2B6	2C19	2C8	2C9	2D6	2E1	3A4
1A2	**271**	61	94	136	86	123	151	93	183
2A6	*2*	**105**	53	50	42	61	58	60	69
2B6	*6*	*4*	**151**	91	67	83	86	65	121
2C19	*3*	*1*	*5*	**218**	90	139	141	58	164
2C8	*2*	*0*	*1*	*0*	**142**	104	81	48	123
2C9	*6*	*1*	*6*	*2*	*2*	**226**	118	72	162
2D6	*11*	*5*	*3*	*6*	*2*	*7*	**270**	76	171
2E1	*5*	*3*	*0*	*4*	*3*	*4*	*3*	**145**	88
3A4	*9*	*2*	*1*	*3*	*2*	*7*	*21*	*6*	**475**

Table 2: Degree of regioselectivity signal overlap
Diagonal: Number of substrates in each set.
Upper-triangle: Number of substrates in both sets with identical observed SOMs.
Lower-triangle: Number of substrates in both sets with identical observed primary SOMs.

Isozyme	1A2	2A6	2B6	2C19	2C8	2C9	2D6	2E1	3A4
1A2	**271**	36	68	98	56	82	102	57	113
2A6	*49*	**105**	36	32	30	42	37	51	42
2B6	*75*	*49*	**151**	62	48	58	64	45	82
2C19	*116*	*40*	*72*	**218**	65	95	107	43	112
2C8	*73*	*34*	*55*	*79*	**142**	79	60	37	73
2C9	*95*	*49*	*66*	*117*	*90*	**226**	79	57	97
2D6	*113*	*48*	*65*	*111*	*65*	*95*	**270**	52	107
2E1	*74*	*53*	*51*	*44*	*42*	*60*	*60*	**145**	51
3A4	*143*	*56*	*98*	*127*	*102*	*119*	*130*	*69*	**475**

Table 3: The percentage of identical regioselectivity signal from the substrates of the column isozyme to be found within the substrates of the row isozyme. Total number of substrates per set is shown on the **diagonal**, save for the merged set which has 680 substrates.

Isozyme	1A2	2A6	2B6	2C19	2C8	2C9	2D6	2E1	3A4	merged
1A2	**271**	34.3	45.0	45.0	39.4	36.3	37.8	39.3	23.8	39.9
2A6	13.3	**105**	23.8	14.7	21.1	18.6	13.7	35.2	8.8	15.4
2B6	25.1	34.3	**151**	28.4	33.8	25.7	23.7	31.0	17.3	22.2
2C19	36.2	30.4	41.1	**218**	45.8	42.0	39.6	29.7	23.6	32.1
2C8	20.7	28.6	31.8	29.8	**142**	35.0	22.2	25.5	15.4	20.9
2C9	30.3	40.0	38.4	43.6	55.6	**226**	29.3	39.3	20.4	33.2
2D6	36.6	35.2	42.4	49.1	42.3	35.0	**270**	35.9	22.5	39.7
2E1	21.0	48.6	29.8	19.7	26.1	25.2	19.3	**145**	10.7	21.3
3A4	41.7	40.0	53.3	51.4	51.4	40.1	39.6	35.2	**475**	69.9

Table 4: CYP-mediated pathway propensities (%) according to the $\frac{\text{\# of oxidized SOMs}}{\text{\# potential SOMs}}$ of the substrates within the Calibration and External sets of 2C9, 2D6 and 3A4.

Pathway	2C9 Cal.	2C9 Ext.	2D6 Cal.	2D6 Ext.	3A4 Cal.	3A4 Ext.
Csp^3 Hydroxylation	21.2	8.9	12.4	5.8	9.9	13.2
Aromatic Hydroxylation	13.0	12.6	10.7	10.7	8.2	14.5
Ring Hydroxylation	11.9	7.1	8.0	11.1	10.4	11.7
O-dealkylation	44.1	50.7	52.1	46.9	26.3	43.4
N-dealkylation	35.7	35.8	27.2	35.2	45.5	42.0
Sulfur(II) Oxidation	53.8	65.2	71.4	50.0	70	40.9
Sulfur(IV) Oxidation	100	0	None	100	100	100
Desulfuration	0	50	100	50	33.3	100
Csp^2 Oxidation	32.7	5.4	0	7.5	9.5	6.8
Aldehyde Oxidation	100	100	None	100	33.3	100
Alcohol Oxidation	28.6	20	0	30.0	6.3	19.4
N-hydroxylation	5.2	6.1	3.4	9.5	3.6	13.8
N-oxide Formation	1.2	3.2	2.6	4.2	4.1	4.7
Nitro-group Reduction	0	0	0	0	0	25.0
Dehalogenation	0	0	0	1.9	1.3	1.7
Uncommon	0.5	0.3	0.1	0.8	0.4	1.1
Overall	11.0	9.7	9.6	9.5	8.6	11.1

Table 5: The percentage of Complete(Com.), Calibration(Cal.) and External(Ext.) 3A4, 2D6 and 2C9 substrate sets with an experimentally observed SOM predicted in the Top-1, -2 or -3 rank-positions by the given method

Feature Set Dataset / Metric	QC	TOP	TOP QC	TOP QC SCR	TOP SCR	SMARTCyp	Stardrop	Schrödinger	Random Model
3A4 Com. Top-1	49.7	61.3	61.5	**67.2**	66.1	62.3	59.7	56.3	11.3
3A4 Com. Top-2	71.4	75.6	77.7	**82.3**	82.1	74.4	74.1	76.4	21.0
3A4 Com. Top-3	79.4	85.1	86.1	**88.6**	88.2	81.4	83.0	84.7	29.6
3A4 Cal. Top-1	46.4	61.7	63.6	**68.2**	66.7	63.9	63.9	59.7	10.2
3A4 Cal. Top-2[a]	72.0	74.8	81.0	**85.7**	81.9	73.1	77.5	80.2	19.4
3A4 Cal. Top-3	81.3	83.5	88.8	**90.7**	88.5	81.1	86.0	88.2	27.6
3A4 Ext. Top-1	49.4	55.8	52.6	57.1	**63.6**	58.8	51.0	49.5	13.5
3A4 Ext. Top-2	62.3	72.1	68.8	72.7	**79.2**	77.2	66.9	68.2	24.5
3A4 Ext. Top-3	76.0	81.2	79.2	84.4	**87.7**	82.0	76.9	77.2	33.7
2D6 Com. Top-1	59.6	67.0	71.1	70.4	**73.0**	46.5	61.5	48.1	11.0
2D6 Com. Top-2	73.7	79.6	83.3	**85.9**	83.7	58.4	75.3	68.1	21.1
2D6 Com. Top-3	80.4	86.3	90.4	**91.1**	90.0	65.2	86.9	82.6	29.9
2D6 Cal. Top-1	61.9	73.9	**73.9**	72.4	72.4	41.4	67.2	43.7	10.6
2D6 Cal. Top-2[b]	75.4	84.3	84.3	**86.6**	85.8	48.5	81.5	66.2	20.2
2D6 Cal. Top-3	85.1	91.0	**91.8**	**91.8**	89.6	56.9	90.1	86.6	28.8
2D6 Ext. Top-1	50.7	58.8	60.3	64.7	**66.9**	51.5	55.9	52.6	11.4
2D6 Ext. Top-2	66.2	75.0	77.2	78.7	**79.4**	68.1	69.2	70.1	21.9
2D6 Ext. Top-3	75.7	81.6	84.6	85.3	**89.0**	73.4	83.6	78.7	31.0
2C9 Com. Top-1	51.8	63.3	63.7	69.5	**72.1**	52.4	61.3	52.0	11.9
2C9 Com. Top-2	73.9	76.5	80.5	**84.5**	84.1	67.3	78.0	72.1	22.2
2C9 Com. Top-3	82.7	87.2	88.5	90.3	**90.7**	77.4	83.3	82.9	31.2
2C9 Cal. Top-1	54.1	64.3	64.3	67.3	**72.4**	53.1	60.7	50.5	12.2
2C9 Cal. Top-2[c]	71.4	76.5	78.6	81.6	**84.7**	67.7	77.4	69.6	22.5
2C9 Cal. Top-3	77.6	83.7	87.8	**89.8**	**89.8**	79.8	83.1	82.9	31.5
2C9 Ext. Top-1	46.9	58.6	56.2	60.2	**64.1**	51.8	61.7	53.1	11.7
2C9 Ext. Top-2	63.3	77.3	78.9	79.7	**80.5**	66.9	78.4	74.0	22.0
2C9 Ext. Top-3	72.7	83.6	86.7	86.7	**87.5**	75.5	83.4	82.8	31.0

a. 77.4 and 61.8 respectively for Merck method and MetaSite Version 2.7.5 for non-updated 3A4 calibration set.
b. 71.9 and 65.4 respectively for Merck method and MetaSite Version 2.7.5 for non-updated 2D6 calibration set.
c. 72.4 and 68.8 respectively for Merck method and MetaSite Version 2.7.5 for non-updated 2C9 calibration set.

Table 6: Lift predictions rates for Complete(Com.), Calibration(Cal.) and External(Ext.) 3A4, 2D6 and 2C9 substrate sets. The Lift metric give each individual substrate a weight value based upon the likelihood of randomly predicting an observed SOM within the Top-1, -2 or -3 rank-positions. Full definition of Lift was given in our prior work.

Feature Set Dataset / Metric	QC	TOP	TOP QC	TOP QC SCR	TOP SCR	SMARTCyp	Stardrop	Schrödinger
3A4 Com. Top-1	45.3	56.7	56.4	**62.3**	60.4	55.8	53.3	51.0
3A4 Com. Top-2	67.7	71.0	73.4	**77.5**	76.2	66.7	69.1	70.5
3A4 Com. Top-3	74.4	80.5	80.7	83.3	**83.5**	74.7	78.7	79.1
3A4 Cal. Top-1	43.3	57.8	58.9	**61.5**	59.7	57.4	57.0	54.1
3A4 Cal. Top-2	68.3	70.1	75.9	**80.1**	75.4	65.9	71.8	73.7
3A4 Cal. Top-3	76.2	78.4	83.3	**84.9**	82.8	74.0	81.2	81.8
3A4 Ext. Top-1	43.4	50.5	47.2	52.4	**61.0**	51.7	43.7	43.3
3A4 Ext. Top-2	57.4	69.1	66.2	70.8	**77.0**	68.7	62.2	62.5
3A4 Ext. Top-3	70.6	78.0	75.6	81.4	**85.4**	76.2	72.6	72.1
2D6 Com. Top-1	57.1	63.2	69.8	67.6	**69.2**	41.1	57.8	43.0
2D6 Com. Top-2	69.8	76.4	81.5	**83.6**	81.6	53.4	71.5	63.3
2D6 Com. Top-3	76.8	84.1	88.2	**89.2**	88.0	59.5	83.6	78.2
2D6 Cal. Top-1	61.4	68.7	**71.6**	71.1	67.1	34.3	64.6	38.1
2D6 Cal. Top-2	72.8	82.2	82.0	**84.2**	83.8	42.7	79.4	60.8
2D6 Cal. Top-3	82.6	89.8	**90.5**	88.8	87.3	49.2	88.3	82.0
2D6 Ext. Top-1	48.4	54.6	56.5	59.0	**64.5**	47.9	50.9	47.9
2D6 Ext. Top-2	61.6	72.4	72.7	75.1	**76.9**	64.2	63.5	65.9
2D6 Ext. Top-3	70.4	77.9	79.9	80.7	**85.3**	69.9	78.9	74.4
2C9 Com. Top-1	49.6	60.1	59.5	66.4	**68.4**	50.0	59.9	45.9
2C9 Com. Top-2	69.4	72.6	75.8	80.2	**80.5**	64.6	76.4	67.7
2C9 Com. Top-3	79.3	84.1	85.4	86.5	**88.8**	74.4	82.1	79.4
2C9 Cal. Top-1	52.9	64.6	62.8	64.5	**72.9**	53.4	61.3	43.8
2C9 Cal. Top-2	68.6	74.1	75.5	78.2	**83.0**	69.0	77.5	66.0
2C9 Cal. Top-3	74.4	81.4	85.5	87.8	**89.2**	78.9	82.6	80.9
2C9 Ext. Top-1	40.3	54.5	50.6	55.0	57.7	47.4	**58.8**	47.5
2C9 Ext. Top-2	58.4	72.0	74.4	73.3	**76.8**	61.2	75.5	69.0
2C9 Ext. Top-3	68.5	78.7	81.9	83.2	**84.7**	71.0	81.6	78.2

Table 7: The percentage of 1A2, 2A6, 2B6, 2C19, 2C8, 2E1 and merged substrate sets with an experimentally observed SOM predicted in the Top-1, -2 or -3 rank-positions by the given method

Feature Set CYP Metric	QC	TOP	TOP QC	TOP QC SCR	TOP SCR	SMARTCyp	Random Model
1A2 Top-1	57.9	64.6	65.3	**70.1**	69.4	62.6	14.3
1A2 Top-2	77.1	77.5	79.7	**83.0**	82.3	78.9	26.0
1A2 Top-3	80.8	88.2	87.1	88.6	**91.9**	85.1	35.6
2A6 Top-1	59.0	62.9	66.7	66.7	**73.3**	69.4	18.3
2A6 Top-2	72.4	81.9	79.0	81.0	**85.7**	83.3	31.9
2A6 Top-3	83.8	86.7	83.8	88.6	**90.5**	87.7	43.5
2B6 Top-1	60.3	62.9	63.6	**64.9**	**64.9**	64.0	14.0
2B6 Top-2	76.8	76.2	80.1	**82.1**	76.8	73.6	24.8
2B6 Top-3	84.8	81.5	86.8	**88.7**	83.4	79.7	33.8
2C19 Top-1	59.6	65.1	67.9	70.6	**72.9**	58.0	10.5
2C19 Top-2	78.4	80.7	82.6	**86.2**	**86.2**	73.7	20.2
2C19 Top-3	87.6	88.1	89.9	**93.6**	92.2	79.9	28.8
2C8 Top-1	57.0	62.7	62.0	67.6	**69.0**	60.4	12.1
2C8 Top-2	77.5	83.1	77.5	**83.8**	**83.8**	73.2	22.6
2C8 Top-3	83.8	88.0	88.0	92.3	**94.4**	81.4	31.7
2E1 Top-1	60.0	62.1	**67.6**	64.8	64.8	62.1	21.1
2E1 Top-2	76.6	79.3	80.7	**82.8**	80.7	81.0	36.5
2E1 Top-3	83.4	85.5	85.5	85.5	**86.2**	84.7	47.7
merged Top-1	56.6	66.3	68.1	70.3	**72.2**	62.0	14.7
merged Top-2	76.0	79.0	81.6	84.1	**86.0**	74.8	26.3
merged Top-3	83.2	86.9	89.7	**90.9**	90.1	81.3	36.0

Table 8: Lift prediction rates for 1A2, 2A6, 2B6, 2C19, 2C8, 2E1 and merged substrate sets

Feature Set CYP Metric	QC	TOP	TOP QC	TOP QC SCR	TOP SCR	SMARTCyp
1A2 Top-1	50.1	60.3	58.8	62.9	**64.4**	58.0
1A2 Top-2	69.5	72.8	75.0	**78.7**	77.1	74.1
1A2 Top-3	74.3	84.6	82.7	84.9	**88.9**	80.5
2A6 Top-1	43.1	48.3	53.2	52.6	**59.9**	57.9
2A6 Top-2	56.8	72.4	67.9	68.5	**75.7**	70.2
2A6 Top-3	73.0	78.3	74.2	81.9	**83.1**	78.8
2B6 Top-1	52.6	56.9	56.0	**57.9**	57.2	57.1
2B6 Top-2	71.2	68.9	74.2	**76.8**	71.0	64.9
2B6 Top-3	80.0	76.0	83.1	**84.4**	79.3	73.9
2C19 Top-1	56.3	62.6	63.7	67.7	**69.8**	54.0
2C19 Top-2	75.0	77.3	80.7	**84.1**	82.7	69.7
2C19 Top-3	85.5	87.2	88.0	**91.9**	89.7	76.0
2C8 Top-1	49.1	60.0	55.1	60.8	**65.9**	56.0
2C8 Top-2	71.9	78.5	71.9	**79.5**	78.1	69.1
2C8 Top-3	80.1	83.2	85.1	87.8	**89.0**	77.7
2E1 Top-1	49.2	52.1	**57.4**	54.4	53.9	47.4
2E1 Top-2	65.1	70.7	69.4	**72.8**	71.3	69.7
2E1 Top-3	72.4	77.5	76.9	77.2	**78.0**	74.9
merged Top-1	49.2	58.9	60.2	62.5	**63.5**	54.3
merged Top-2	68.1	72.9	76.0	77.3	**79.5**	66.6
merged Top-3	76.4	81.7	84.5	**85.8**	85.0	73.9

Table 9: Robustness of 2C9, 2D6 and 3A4 Models: Performance differences between External and Calibration prediction rates for Standard and Lift metrics

Feature Set Dataset / Metric	QC	TOP	TOP QC	TOP QC SCR	TOP SCR	SMARTCyp	Stardrop	Schrödinger	Random
3A4 Standard Top-1	2.9	-5.8	-11.0	-11.1	-3.0	-5.2	-12.9	-10.5	2.3
3A4 Standard Top-2	-9.6	-2.7	-12.2	-12.9	-2.7	4.0	-10.6	-12.0	5.1
3A4 Standard Top-3	-5.3	-2.3	-9.6	-6.2	-0.8	0.9	-9.2	-11.0	6.1
3A4 Lift Top-1	0.1	-7.3	-11.7	-9.1	1.2	-5.7	-13.3	-9.8	
3A4 Lift Top-2	-10.9	-1.0	-9.7	-9.3	1.6	2.7	-9.6	-11.2	
3A4 Lift Top-3	-5.6	-0.4	-7.6	-3.4	2.6	2.2	-8.6	-8.3	
2D6 Standard Top-1	-11.2	-15.1	-13.6	-7.7	-5.5	10.1	-11.3	8.9	0.8
2D6 Standard Top-2	-9.2	-9.3	-7.1	-7.9	-6.4	19.6	-12.3	3.9	1.7
2D6 Standard Top-3	-9.3	-9.4	-7.2	-6.5	-0.6	16.5	-6.5	-7.9	3.2
2D6 Lift Top-1	-13.0	-14.1	-15.1	-12.1	-2.6	13.6	-13.7	9.8	
2D6 Lift Top-2	-11.3	-9.8	-9.2	-9.1	-6.9	21.5	-15.9	5.1	
2D6 Lift Top-3	-12.1	-11.9	-10.6	-8.2	-2.0	20.8	-9.4	-7.6	
2C9 Standard Top-1	-7.2	-5.7	-8.0	-7.2	-8.4	-1.2	1.0	-2.6	-0.5
2C9 Standard Top-2	-8.1	0.8	0.3	-1.9	-4.2	-0.8	0.9	-5.4	-0.5
2C9 Standard Top-3	-4.9	-0.1	-1.0	-3.1	-2.3	-4.3	0.3	0.1	-0.5
2C9 Lift Top-1	-12.6	-10.1	-12.2	-9.5	-15.2	-6.1	-2.7	3.7	
2C9 Lift Top-2	-10.2	-2.1	-1.1	-5.0	-6.2	-7.8	-2.0	3.0	
2C9 Lift Top-3	-5.9	-2.6	-3.6	-4.7	-4.5	-7.9	-1.0	-2.7	

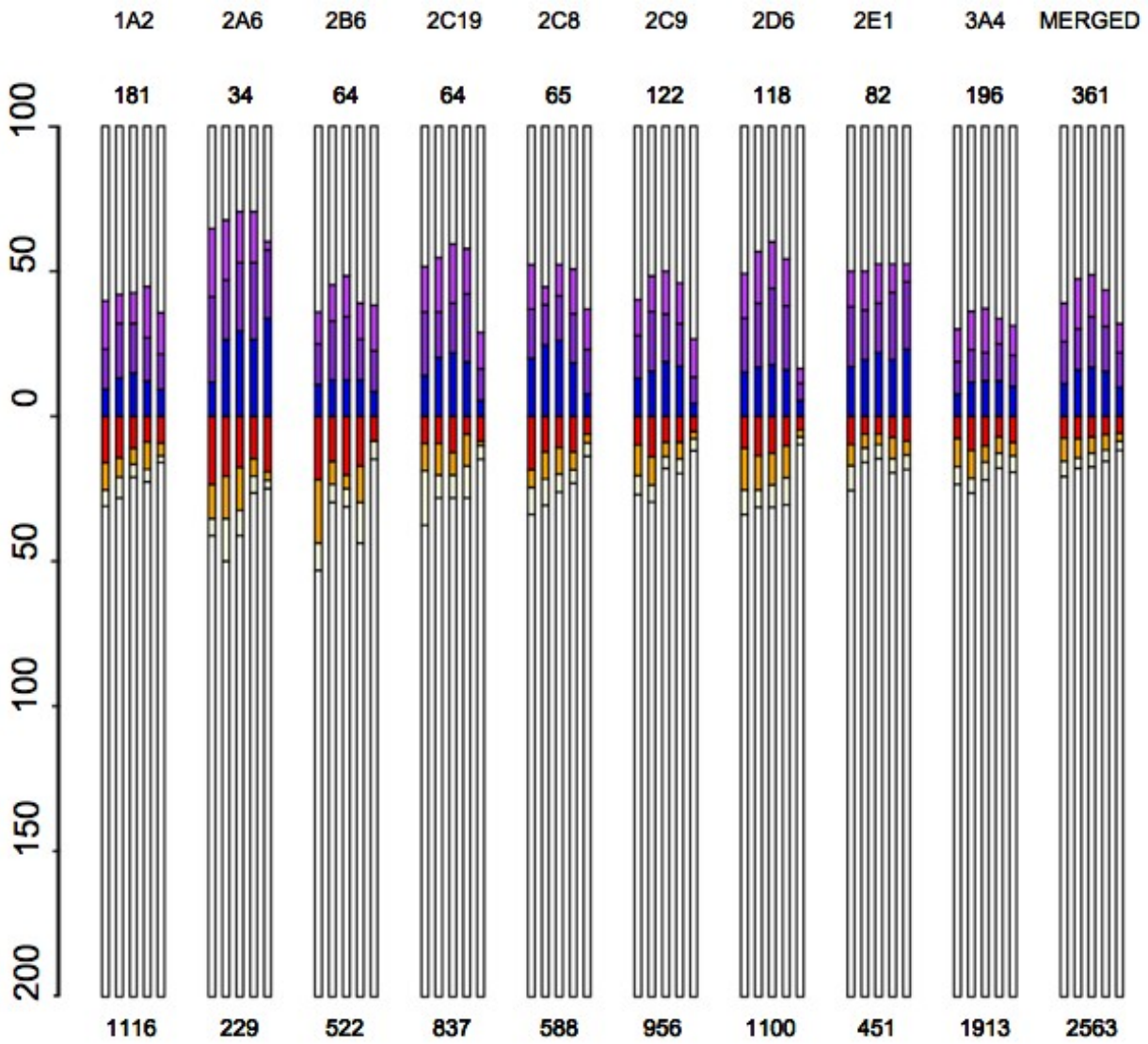

Figure 1: Aromatic ring hydroxylation true-positive and false-negative rates of (from left to right) TOP, TOP QC, TOP QC SCR, TOP SCR and SMARTCyp models for all substrate sets. For the majority of models QC modes have improved TP rates relative to non-QC models.

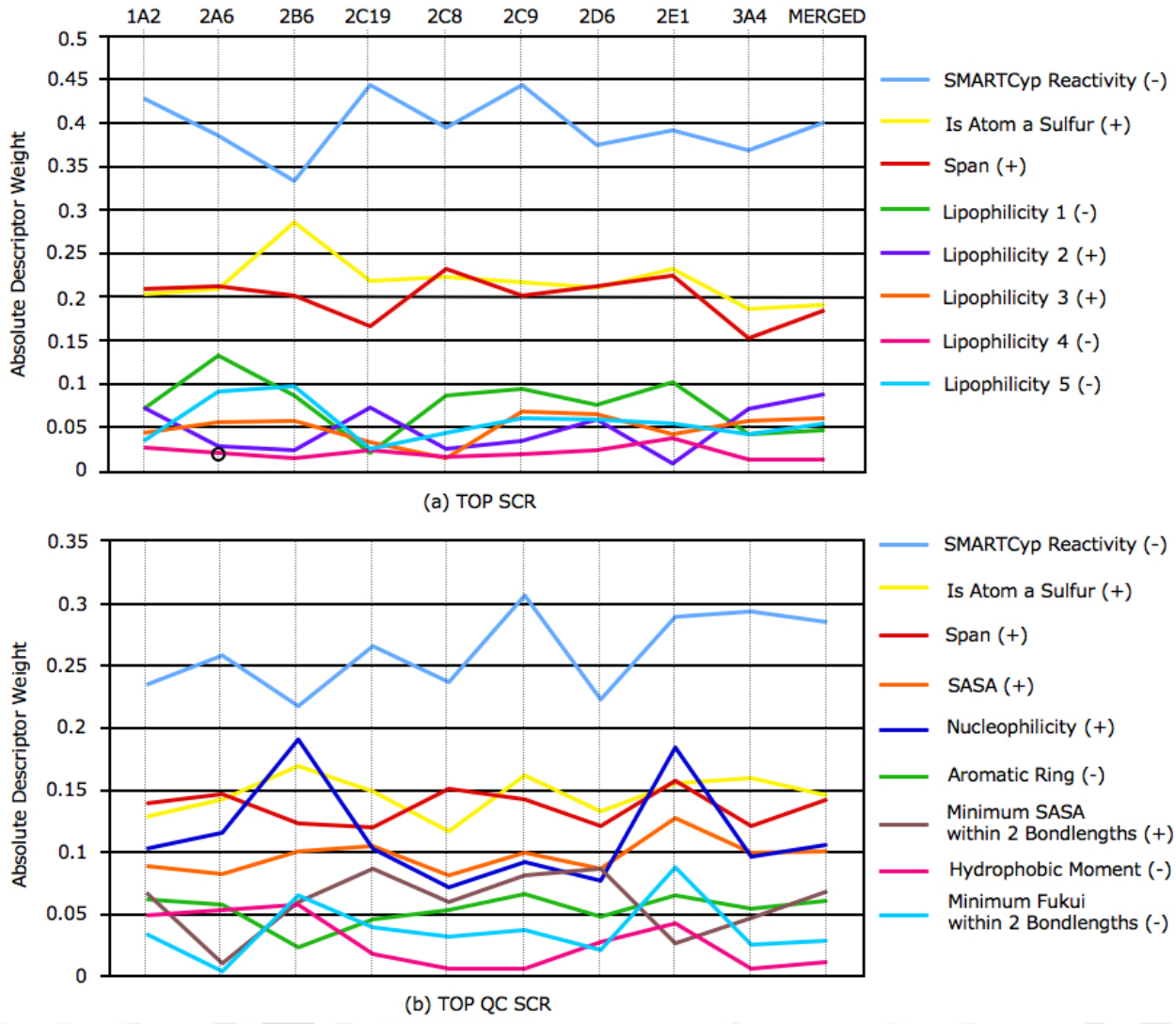

Figure 2: The average absolute weight values of influential descriptors for each isozyme model. The positive or negative correlation of each descriptor with observed regioselectivity is designated on the right of each label. The one exception is the correlation of lipophilicity 4 within the TOP SCR model of 2A6, designated by a black cirlce.

Table 10: The 10 descriptors with the highest absolute weights for each TOP SCR and TOP QC SCR isozyme model

1A2

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
Span(+)	Span(+)
NA_0_S(+)	NA_0_S(+)
NA_3_N(-)	nucleophilicity(+)
BASE_GROUP_NON_HYD_BOND_COUNT(+)	atom_area(+)
NA_1_C(-)	atom_area_D2_min(+)
LLipo_2(+)	BASE_GROUP_In_AromaticRing(-)
BASE_GROUP_In_AromaticRing(-)	LLipo_1(-)
LLipo_1(-)	Hydrophobic_Moment(-)
PA_2_H(-)	self_charge(+)

2A6

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
Span(+)	Span(+)
NA_0_S(+)	NA_0_S(+)
LLipo_1(-)	nucleophilicity(+)
NA_0_O(-)	atom_area(+)
BASE_GROUP_ROTATABLE_BONDS_2(+)	LLipo_1(-)
LLipo_5(-)	NA_0_O(-)
PA_1_N(-)	BASE_GROUP_ROTATABLE_BONDS_2(+)
PA_1_C(+)	fukui_D1_norm(-)
PA_4_C(+)	self_charge_D2_min(-)

2B6

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
NA_0_S(+)	nucleophilicity(+)
Span(+)	NA_0_S(+)
NA_0_N(-)	Span(+)
LLipo_5(-)	atom_area(+)
NA_0_O(-)	pi_pi_bo_D1_BS_min(+)
LLipo_1(-)	NA_0_N(-)
NA_4_Hp(-)	fukui_D2_min(-)
NA_0_A(-)	atom_area_D2_min(+)
PA_3_C(+)	NA_0_A(-)

2C19

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
NA_0_S(+)	NA_0_S(+)
Span(+)	Span(+)
NA_4_Hp(-)	atom_area(+)
BASE_GROUP_In_MultipleRings(-)	nucleophilicity(+)
NA_2_S(+)	atom_area_D2_min(+)
NA_1_C(-)	LLipo_1(-)
NA_2_Hp(-)	BASE_GROUP_In_MultipleRings(-)
NA_0_N(-)	NA_4_S(-)
NA_4_C(-)	self_charge(+)

2C8

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
Span(+)	Span(+)
NA_0_S(+)	NA_0_S(+)
NA_2_D(-)	atom_area(+)
BASE_GROUP_In_MultipleRings(-)	nucleophilicity(+)
NA_4_Po(-)	LLipo_1(-)
PA_2_D(-)	BASE_GROUP_In_MultipleRings(-)
LLipo_1(-)	NA_2_D(-)
PA_4_Po(-)	atom_area_D2_min(+)
NA_4_S(-)	NA_4_Po(-)

2C9

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
NA_0_S(+)	NA_0_S(+)
Span(+)	Span(+)
LLipo_1(-)	atom_area(+)
NA_4_S(-)	nucleophilicity(+)
NA_4_Po(-)	atom_area_D2_min(+)
NA_4_C(-)	LLipo_1(-)
NA_0_O(-)	LLipo_5(-)
NA_2_D(-)	LLipo_3(+)
BASE_GROUP_In_Ring3(-)	BASE_GROUP_In_AromaticRing(-)

2D6

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
Span(+)	NA_0_S(+)
NA_0_S(+)	Span(+)
PA_3_C(+)	atom_area_D2_min(+)
NA_4_N(-)	atom_area(+)
NA_4_Hp(-)	nucleophilicity(+)
BASE_GROUP_In_MultipleRings(-)	BASE_GROUP_ROTATABLE_BONDS_2(+)
LLipo_1(-)	PA_3_N(-)
BASE_GROUP_ROTATABLE_BONDS_2(+)	BASE_GROUP_In_AromaticRing(-)
NA_3_Hp(-)	PA_3_C(+)

2E1

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
NA_0_S(+)	nucleophilicity(+)
Span(+)	Span(+)
BASE_GROUP_In_MultipleRings(+)	NA_0_S(+)
NA_3_H	atom_area
LLipo_1(-)	fukui_D2_max(+)
PA_4_Po(-)	fukui_D2_min(-)
BASE_GROUP_ROTATABLE_BONDS_3(-)	BASE_GROUP_In_MultipleRings(+)
NA_3_X(+)	fukui_D2_norm(+)
PA_1_A(-)	sig_pi_mulliken_D1_BS_max(+)

3A4

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
NA_0_S(+)	NA_0_S(+)
Span(+)	Span(+)
NA_0_O(-)	atom_area(+)
BASE_GROUP_In_AromaticRing(-)	nucleophilicity(+)
LLipo_2(+)	LLipo_1(-)
LLipo_3(+)	BASE_GROUP_In_AromaticRing(-)
NA_1_C(-)	NA_0_O(-)
BASE_GROUP_In_MultipleRings(-)	BASE_GROUP_In_MultipleRings(-)
BASE_GROUP_In_Ring7(+)	atom_area_D2_min(+)

merged

TOP SCR	TOP QC SCR
SMARTCyp_reactivity(-)	SMARTCyp_reactivity(-)
NA_0_S(+)	NA_0_S(+)
Span(+)	Span(+)
LLipo_2(+)	nucleophilicity(+)
BASE_GROUP_In_AromaticRing(-)	atom_area(+)
NA_1_C(-)	atom_area_D2_min(+)
NA_0_O(-)	BASE_GROUP_In_AromaticRing(-)
NA_4_Hp(-)	LLipo_1(-)
LLipo_3(+)	LLipo_3(+)
NA_4_C(-)	NA_0_O(-)

Model Stability

An individual MIRank model is a vector of numbers representing the relative importance of each descriptor towards the ranking of the observed SOMs of each substrate over the non-observed SOMs of the same substrate for all compounds upon which the model is trained. The weight of each descriptor may be positive or negative, indicating the correlation of values of the descriptor with observed regioselectivity trends. The absolute weight values of the 149 TOP SCR descriptors of all ten substrate set models are shown in decreasing order in Figure 3a. While there are slight differences between models, they all share the general trend of having a minority(10-31) of descriptors with high weight values (>.05) and a large majority of descriptors with lower weight values (<.05) that gradually asymptote to zero. A view of the 50 descriptors with the highest absolute weight for TOP, TOP QC, TOP SCR and TOP QC SCR models of the merged set is given in Figure 3b. Close examination reveals that the highest weighted descriptor of the SCR models, which is in fact the SMARTCyp reactivity descriptor, has a higher weight value than the top weighted descriptor of non-SCR models. One of the consequences of this is that the weights of all other descriptors within these models have lower weights than they otherwise would have had. This is expected to happen whenever new descriptors that contain additional signal are added to model, as MIRank elucidates the maximum amount of complementary information contained within a given descriptor set. Consequently the absolute weight values of QC models, which have 392 additional descriptors, fall slightly lower than non-QC values for the first 70 or so descriptors and take longer to asymptote, as illustrated in Figure 3c. The question of where and when QC signal is valid is then synonymous with how relevant the asymptotic area for the lowest ($541 - \sim 70$) weighted descriptors is for a particular substrate. As an additional complication, each prediction made by RS-Predictor is in fact the aggregation of the individual substrate predictions of 10 independently generated MIRank models.

RS-Predictor operates by employing 10-fold cross-validation, essentially dividing a training set into ten partitions and making a prediction for the substrates within each partition based upon the elucidated signal from the nine other partitions. However to ensure that results are unbiased

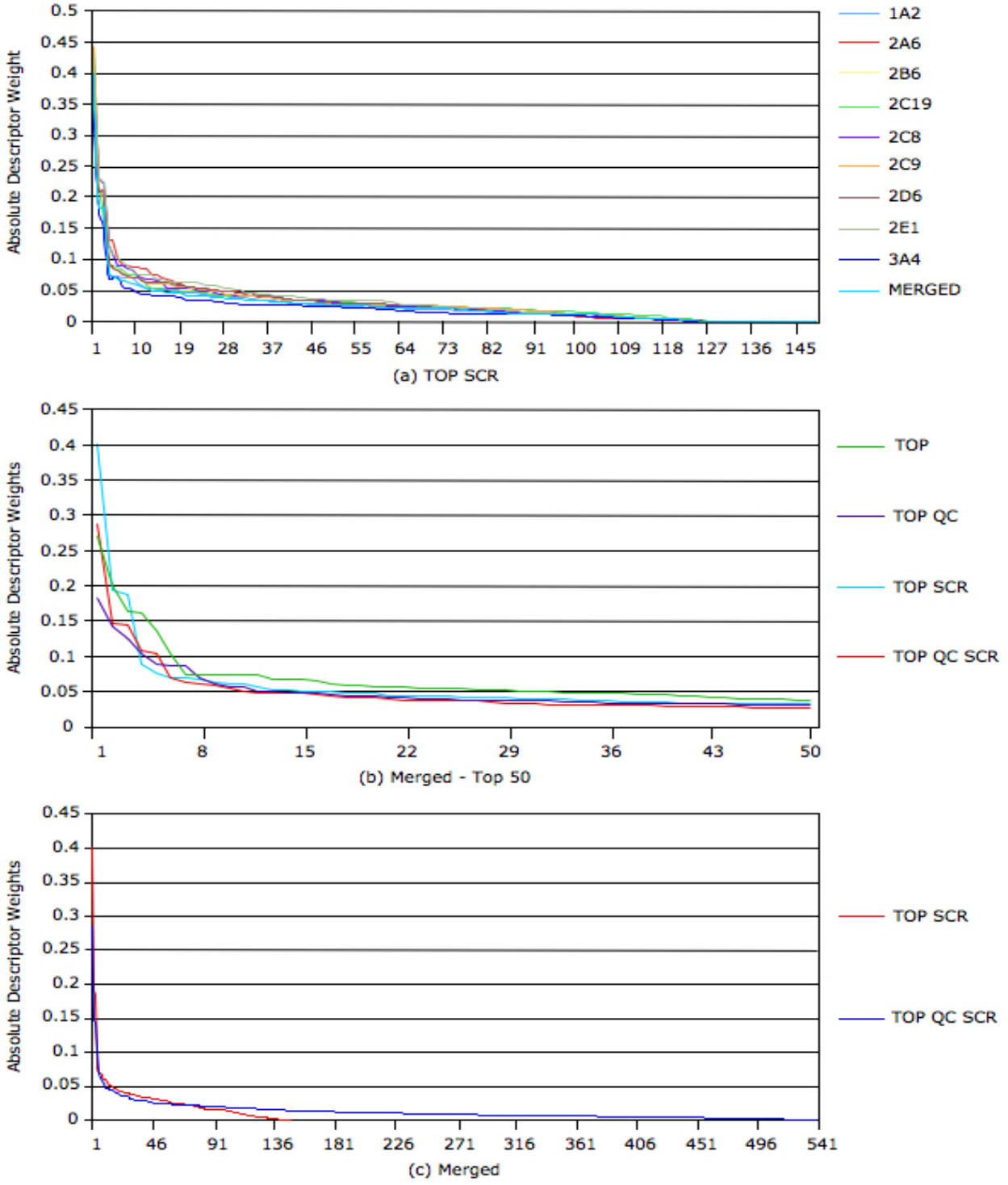

Figure 3: (a) Absolute weights of the 149 TOP SCR descriptors for each substrate set model sorted by decreasing value. (b) The values of the 50 descriptors with the highest absolute weights for TOP(148), TOP QC(540), TOP SCR(149) TOP QC SCR(541) models of the merged substrate set. (c) Absolute weights of TOP SCR and TOP QC SCR descriptors for the merged substrate set.

from the initial divide, this process is repeated ten times, giving each individual substrate ten different SOM rank-orderings that were obtained from 90% of the substrates in the given set. These independent SOM rank-orderings are then merged into a single consensus rank-ordering using techniques of rank aggregation to ensure that end-user medicinal chemists will only have to look at a single optimized prediction for each predicted substrate. Another benefit of this technique is that the accuracy of the aggregated prediction tends to be greater than the average accuracy of each individual set of cross-validated predictions. We were the first group to incorporate the concept of rank aggregation with regioselectivity modeling, but in our initial work RS-Predictor was only applied to CYP 3A4 using TOP QC descriptors. Further investigation into the utility of rank aggregation, as well as the relative stability of predictions made by independent training set partitions, are made in Figure 4. It is important to note that the weight values used earlier in Figure 3 were averaged across the ten different models for the particular isozyme, and as such have corresponding standard deviations.

The most important differentiation between TOP SCR and TOP QC SCR models is that the performances of individual QC models are much less stable. TOP QC SCR models have visibly larger ranges in overall cross-validated prediction quality, and their average rates are lower than the average rates of non-QC models for all isozymes save 2B6. This corroborates previously mentioned views that semiempirical-derived models should be interpreted and used cautiously for compounds outside of the initial training set (*Mayeno et. al 2009*). The likely cause of this problem is that the much larger asymptotic QC signal elucidated from 90% of training set substrates is valid for the remaining test partition in some cases, and not valid in others. What is very interesting however is that rank aggregation, essentially the merging of independently predicted signal identified from multiple training set partitions, gives significantly greater performance increases for QC models than they do for non-QC models. This makes sense, as the greater the variation between individual predictions, the more likely overlaps between those predictions stem from a successful application of encoded regioselectivity signal, as opposed to noise dependent upon a particular training set partition. The use of rank aggregation upon multiple cross-validated predictions therefore cir-

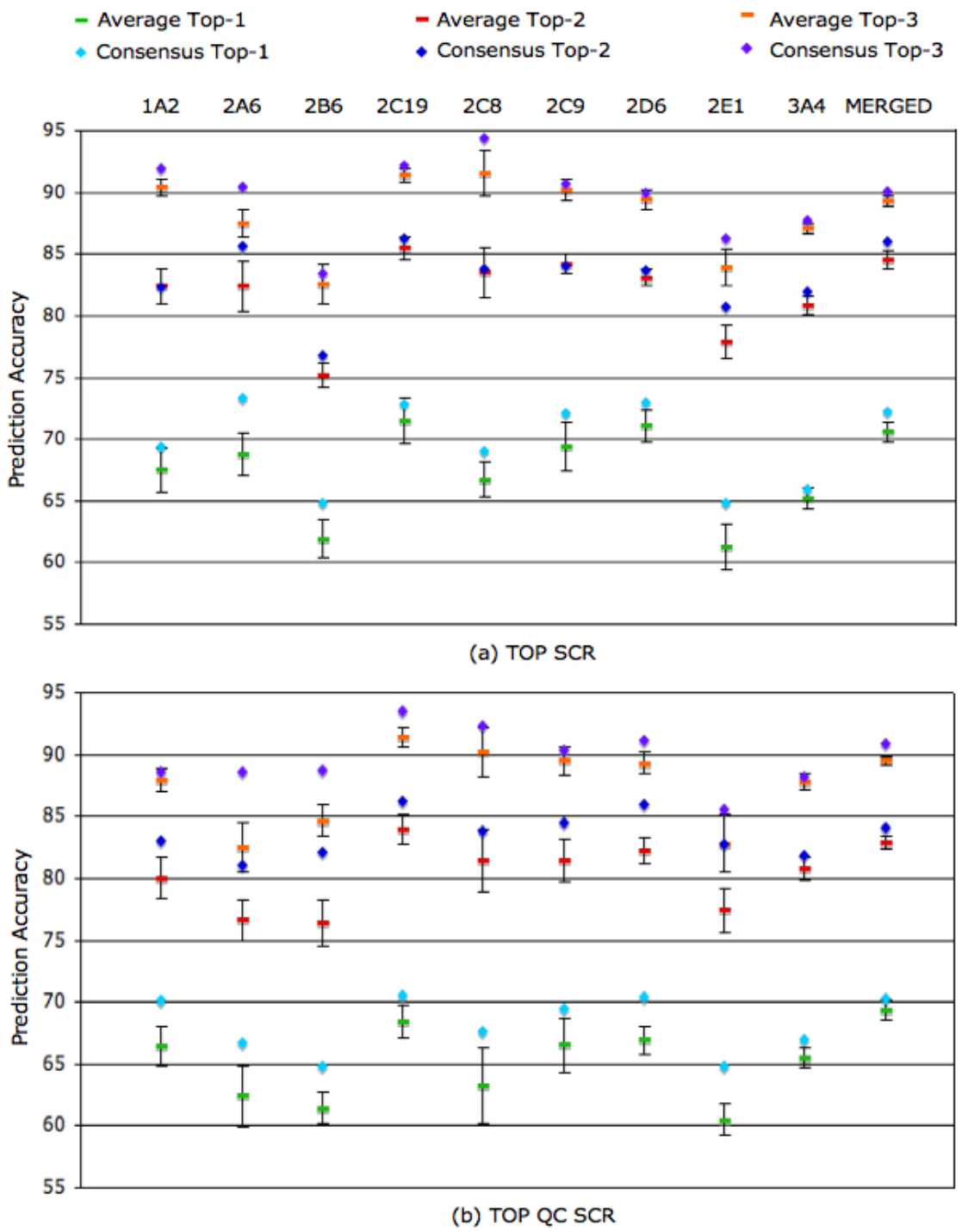

Figure 4: The percentage of each substrate set with an observed SOM predicted within the Top-1,-2 or -3 rank-positions by one of ten individual cross-validated models, or by the rank aggregated consensus of the predictions made by those models.

cumvents the difficulties associated with overtraining models through the use of large number of descriptors, which is likely why TOP QC SCR models have optimal prediction rates for 8 of the

10 reported substrate sets.