Electronic Supplementary Material (ESI) for CrystEngComm.

The *CSD* and knowledge databases: from answers to questions

Alexander P. Shevchenko, [abc] Roman A. Eremin, [ab] Vladislav A. Blatov [*ab]

Samara Center for Theoretical Materials Science (SCTMS), Samara University, Ac. Pavlov St. 1, 443011 Samara, Russian Federation.

Samara Center for Theoretical Materials Science (SCTMS), Samara State Technical University, Molodogvardeyskaya St. 244, 443100 Samara, Russian Federation.

Samara Branch of P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Novo-Sadovaya St. 221, 443011 Samara, Russian Federation.

Supporting Information

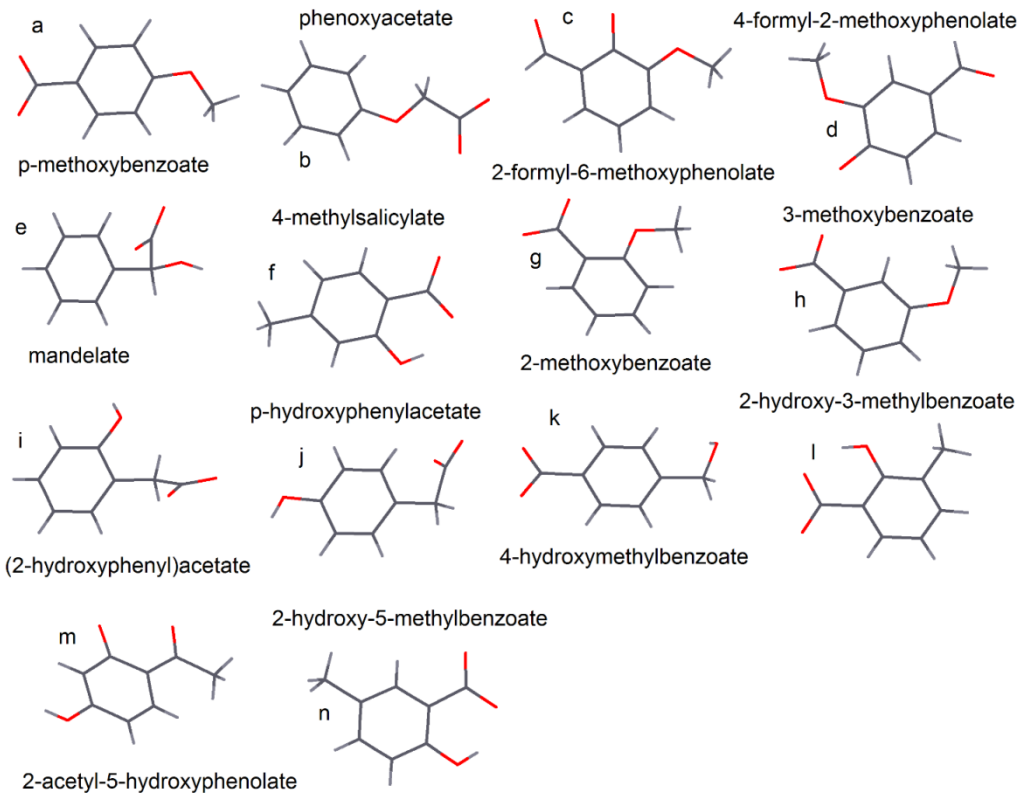

Figure S1. Chemical formulae of the isomers of the $C_8H_7O_3$ composition.

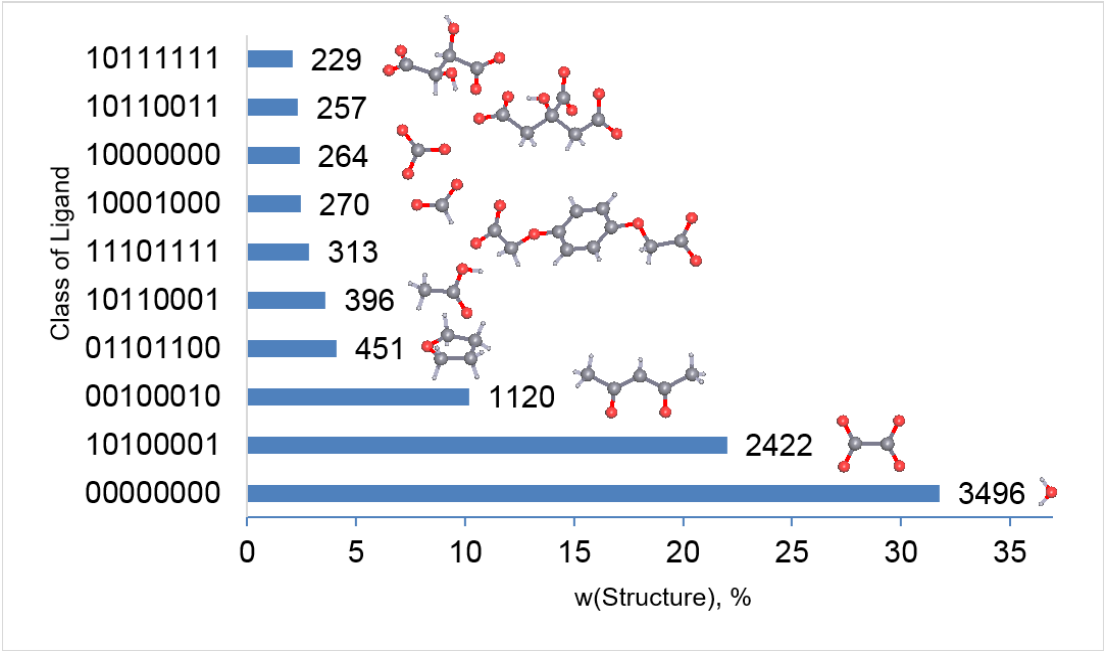

Figure S2. Top ten most abundant topological fingerprints (classes of ligands).

Table S3. Distribution of the first ten most abundant coordination figures of ligands on coordination numbers (CN) in the range 1–4 in the underlying nets. The first three coordination figures for each coordination number are shown in Fig. 4.

Coordination figure	N	ω,%	Example (ligand name, formula, CSD RefCode)
CN=1	**4253**	**31.7**	
L-1	4253	100	heptane-3,5-dionato, $C_7H_{11}O_2^-$, CIKSAT
CN=2	**3602**	**26.8**	
A-2{60}	622	17.3	acetate, $C_2H_3O_2^-$, CUAQAC01
A-2{45}	557	15.5	3-carboxyphenoxyacetato, $C_9H_7O_5^-$, JAQPUN
A-2{90}	472	13.1	aqua, H_2O, OQOGUY
L-2 = A-2{180}	420	11.7	terephthalate, $C_8H_4O_4^{-2}$, DIKQET04
A-2{75}	376	10.4	mandelate, $C_8H_6O_3^-$, INIROM
A-2{105}	351	9.7	aqua, H_2O, XADLIX
A-2{120}	205	5.7	5-carboxybenzene-1,3-dicarboxylato, $C_9H_4O_6^{-2}$, VUXBUL01
A-2{135}	179	5.0	hydroxyacetate, $C_2H_3O_3^-$, GDHOAC11
A-2{30}	156	4.3	benzoate, $C_7H_5O_2^-$, NEXZAS
A-2{150}	131	3.6	4-hydroxyphenoxo, $C_6H_4O_2^{-2}$, NIVFUT
Other 3	*133*	*3.7*	
CN=3	**1950**	**14.5**	
FAN-3{60,75,135}	72	3.7	formate, CHO_2^-, NIHBEM
TPY-3{90,105,120}	63	3.2	2-hydroxybutane-1,4-dioate, $C_4H_4O_5^{-2}$, FAYXEJ02
TP-3{105,120,135}	62	3.2	malonate, $C_3H_2O_4^{-2}$, MALMND01
FAN-3{60^2,120}	61	3.1	cyclobutane-1,1-dicarboxylate, $C_6H_6O_4^{-2}$, JUPJIP
TPY-3{90,105^2}	59	3.0	hydroxo, OH^-, UWAHEH
TPY-3{45,135,165}	59	3.0	benzene-1,3-dioxydiacetate, $C_{10}H_8O_6^{4-}$, BIPJAM
TPY-3{45,60,90}	57	2.9	furan-3-carboxylate, $C_5H_3O_3^-$, DESKIW
FAN-3{45^2,90}	54	2.8	9H-xanthene-9-carboxylate, $C_{14}H_9O_3^-$, NIDDEL
TP-3{75,135,150}	52	2.7	formate, CHO_2^-, LOSKUA
TPY-3{45^3}	50	2.6	phenoxo, $C_6H_5O^-$, JODDAH01
Other 45	*1361*	*69.8*	
CN=4	**1754**	**13.1**	
RAP-4{45^2,135^2,180^2}	115	6.6	3,3'-dimethoxybiphenyl-4,4'-dicarboxylate, $C_{16}H_{12}O_6^{-2}$, WEDNUQ
SPY-4{45,60,75,90,120^2}	108	6.2	benzene-1,2-dicarboxylate, $C_8H_4O_4^{-2}$, LEJPIA
RAP-4{60^2,120^2,180^2}	106	6.0	succinate, $C_4H_4O_4^{-2}$, MAQZIP
RAP-4{30^2,135,150,180^2}	94	5.4	terephthalate, $C_8H_4O_4^{-2}$, IXODUV
RAP-4{75^2,105^2,180^2}	71	4.0	butane-1,2,3,4-tetracarboxylate, $C_8H_6O_8^{-4}$, IZEGEA
SS-4{75,90,105,120^2,150}	69	3.9	tartarate, $C_4H_4O_6^{-2}$, ZOMREA
RAP-4{60^2,90,120,135,150}	59	3.4	succinate, $C_4H_4O_4^{-2}$, IHILUI
FAN-4{60,75^2,135^2,150}	57	3.2	furan-2,5-dicarboxylato, $C_6H_2O_5^{-2}$, VAZLAL
SP-4{90^4,180^2}	56	3.2	cyclobutane-1,2,3,4-tetrone, $C_4O_4^{-2}$, RISXAU
T-4{90^2,105,120^2,135}	51	2.9	2,3-dihydroxybutanedioate, $C_4H_4O_6^{-2}$, JIFXIG
Other 47	*968*	*55.2*	
CN>4	**1877**	**14.0**	
Total	**13436**		

Table S4. Distribution of the first ten most abundant topological types (TT) of underlying nets of MOFs in the standard representation for different dimensionalities (see Fig. 6, a-l from left to right)

TT	N	ω,%	Example				
			CSD RefCode	Ligand	Metal	Not coordinated	Figure
0D	**2087**	**33.8**					
2M4-1	323	15.5	HEXVOY	$C_{10}H_{11}O_5^-\|C_{15}H_{11}O_2^-$	Dy	-	a
1,4M5-1	310	14.9	CEVNEZ	$C_{13}H_7O_2^-\|C_4H_8O$	Ni	C_4H_8O	b
1,3M4-1	222	10.6	ACACGA03	$C_5H_7O_2^-$	Ga	-	c
1,2M3-1	176	8.4	CUBEAC02	$C_{10}H_9O_2^-$	Cu	-	
1,6M7-1	172	8.2	GETDIV	$C_9H_4O_6^{-2}\|H_2O$	Mg	H_2O	
2,4M6-3	132	6.3	MEQPAC	$C_{28}H_{21}O_2^-\|C_4H_{10}O$	Rh	$C_4H_{10}O$	
3,4M6-1	78	3.7	WEXBUZ	$C_8H_{13}O_3^-$	Li|Ni	-	
1,5M6-1	73	3.5	CIKSAT	$C_7H_{11}O_2^-\|H_2O$	Yb	-	
3M8-1	34	1.6	RIGNAZ	$C_{14}H_{20}O_2^-\|C_4H_8O$	Na	-	
2,3M5-1	31	1.5	RIGNED	$C_{14}H_{20}O_2^-\|C_4H_8O$	Li	C_4H_8O	
Other 169	*536*	*25.7*					
1D	**903**	**14.6**					
2C1	280	31.0	FOZHUX12	$C_2O_4^{-2}\|H_2O$	Mn	-	d
(4,4)(0,2)	136	15.1	BOVQAF02	$C_8H_7O_2^-\|H_2O$	Pb	-	e
2,4C4	120	13.3	BADRAZ01	$CHO_2^-\|HO^-$	Cu	-	f
2,6C1	94	10.4	CEXBOZ	$C_9H_9O_2^-$	La	-	
3,4C4	40	4.4	AFEHOK01	$C_8H_7O_2^-$	Cu	-	
2,2,5,6C1	21	2.3	VORHOB	$C_2H_3O_2^-\|H_2O$	Co|Gd	H_2O	
3,5C2	14	1.6	CATPAL04	$C_8H_4O_4^{-2}\|H_2O$	Ca	-	
2,2,3C6	10	1.1	HORTIS	$C_{14}H_4O_8^{-4}\|H_2O$	Ni	H_2O	
3,4C3	10	1.1	DAYZUB	$C_4H_5O_2^-\|O^{-2}\|H_2O$	Na|U	-	
3,4,5C2	10	1.1	DEHGEF	$C_6H_8O_4^{-2}\|C_8H_4O_5^{-2}\|H_2O$	Pr	H_2O	
Other 87	*168*	*18.6*					
2D	**1036**	**16.8**					
sql	201	19.4	CAPHTH03	$C_8H_4O_4^{-2}\|H_2O$	Ca	-	g
hcb	114	11.0	GUQXIZ06	$C_9H_4O_6^{-2}\|H_2O$	Cd	-	h
fes	58	5.6	OWEZEW01	$C_9H_4O_6^{-2}\|H_2O$	Ca	-	i
bey	51	4.9	CETHER	$C_6H_8O_4^{-2}\|C_8H_4O_4^{-2}\|H_2O$	Gd	-	
kgd	40	3.9	PAXMOT	$C_{11}H_{21}O_2^-$	Pb	-	
bex	32	3.1	WEWHEO	$C_{16}H_{12}O_6^{-2}\|H_2O$	Na	-	
(6,3)Ia	27	2.6	XEQKEM	$C_{10}H_3O_8^{-3}\|H_2O$	La	-	
3,5L60	18	1.7	DAQRUK	$C_4H_4O_6^{-2}\|C_8H_4O_4^{-2}\|H_2O$	Na|Sm	-	
(6,3)IIa	15	1.4	IGIVIF	$C_{12}H_6O_4^{-2}$	Cd	-	
5,5L4	15	1.4	YAMFAW	$C_{10}H_{14}O_4^{-2}$	Pb	-	
Other 253	*465*	*44.9*					
3D	**2145**	**34.8**					
pcu	84	3.9	ETAFOW06	CO_3^{-2}	Fe	-	j
nia	50	2.3	UDOREM02	CHO_3^-	Na	-	k
pts	45	2.1	RUZBUL	$C_{16}H_{12}O_6^{-2}$	Zn	-	l
bnn	41	1.9	YETQOG	$C_9H_5O_6^-\|H_2O$	Na	-	
dia	35	1.6	ESADUZ	$C_{14}H_8O_5^{-2}\|H_2O$	Cd	-	
flu	31	1.4	NEZQAN	$C_7H_4O_3^{-2}\|CH_4O$	Li	-	
3,6,6T1	29	1.4	ZZZSDW03	$C_2H_3O_2^-$	Mg	-	
sra	24	1.1	CARGEK	$C_9H_3O_6^{-3}\|H_2O$	Bi	H_2O	
tcs	18	0.8	MEJPEZ	$C_2O_4^{-2}\|C_8H_4O_5^{-2}\|H_2O$	Nd	H_2O	
mab	18	0.8	NIFORM02	$CHO_2^-\|H_2O$	Ni	-	
Other 875	*1770*	*82.5*					
Total	**6171**						

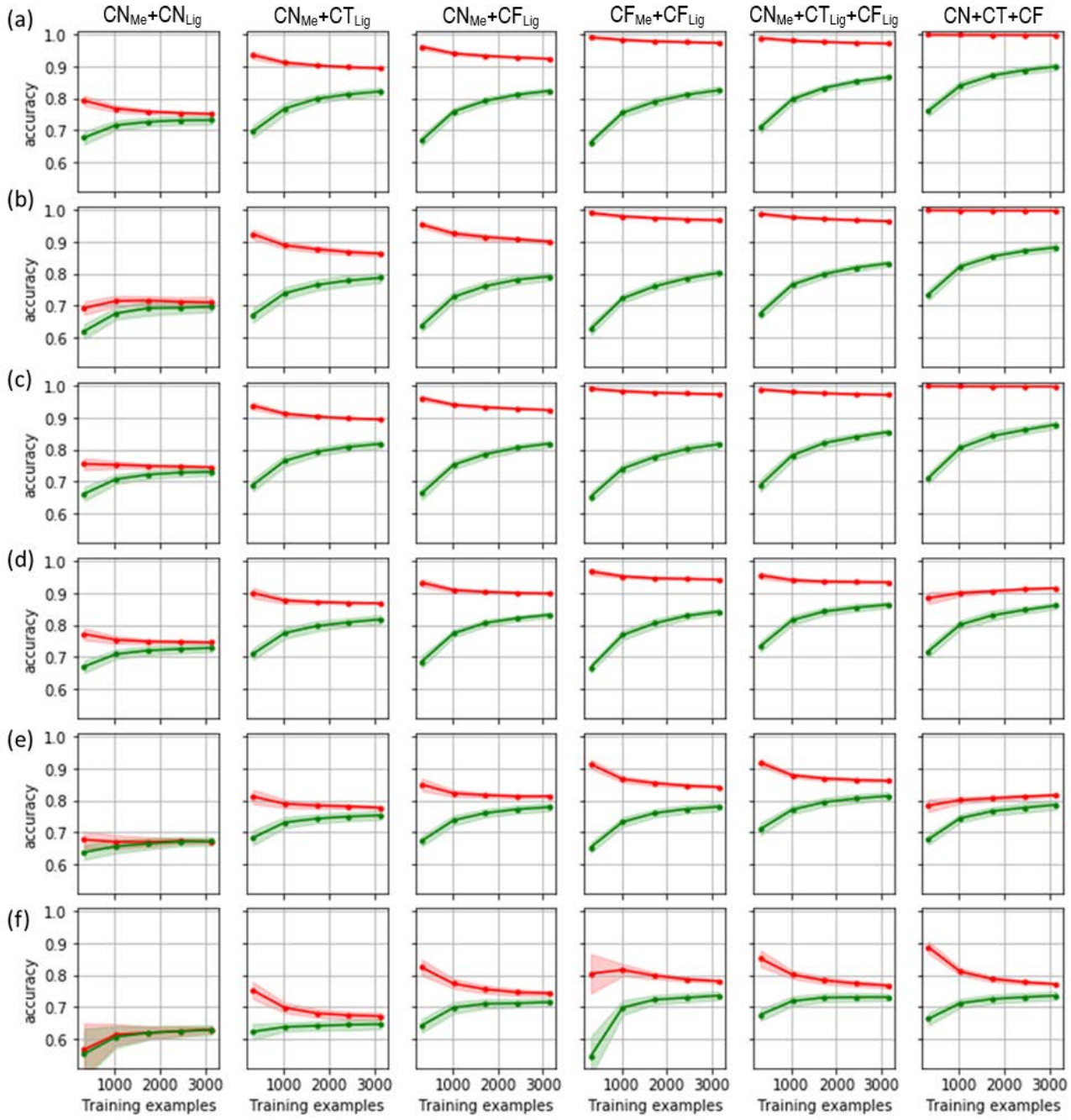

Figure S3. Cross-validation-based learning curves for prediction of dimensionalities of coordination networks for the trained (a) random forest classifier, (b) KNeighbors classifier, (c) decision tree classifier, (d) SVC, (e) logistic regression classifier and (f) Gaussian NB classifier obtained for 50 independent stratified shuffled splits of the full dataset (training set – 80%, testing set – 20%) and 5 different sizes of training subsets chosen from the training set. Red and green lines represent dependencies of the classification accuracies of the trained models on the training and testing datasets, respectively. Translucent areas along the lines correspond to one standard deviation of the obtained scores.

Table S5. Cross-validation-based accuracies for the testing data (20% of the full dataset) corresponding to the maximum training set size available (80% of the full dataset). Feature set corresponds to a certain set of geometrical-topological descriptors (see text for more details). Accuracies obtained are given with their standard deviations based on the 50 independent train/test splits of the full dataset.

Classifier/ Feature set	RF	KNN	CART	SVC	LR	GNB
CN_{Me}+CN_{Lig}	0.732 ± 0.013	0.698 ± 0.017	0.731 ± 0.012	0.729 ± 0.013	0.673 ± 0.013	0.629 ± 0.016
CN_{Me}+CT_{Lig}	0.822 ± 0.013	0.788 ± 0.016	0.818 ± 0.012	0.818 ± 0.014	0.754 ± 0.014	0.647 ± 0.016
CN_{Me}+CF_{Lig}	0.824 ± 0.010	0.792 ± 0.013	0.819 ± 0.011	0.832 ± 0.009	0.779 ± 0.013	0.716 ± 0.013
CF_{Me}+CF_{Lig}	0.826 ± 0.013	0.803 ± 0.012	0.817 ± 0.012	0.842 ± 0.010	0.780 ± 0.013	0.736 ± 0.013
CN_{Me}+CT_{Lig}+CF_{Lig}	0.866 ± 0.009	0.833 ± 0.009	0.855 ± 0.010	0.864 ± 0.010	0.814 ± 0.013	0.732 ± 0.011
CN+CT+CF weighted scheme	0.900 ± 0.011	0.883 ± 0.012	0.879 ± 0.012	0.861 ± 0.010	0.787 ± 0.015	0.736 ± 0.013

Table S6. Predicted probabilities of the underlying topology of the coordination polymers composed by bivalent metals and formate, acetate, oxalate, or terephtalate ligands depending on the coordination formula and shape of coordination figure. The best CN+CT+CF weighted scheme in the Random Forest Classifier is used.

CF of Ligands	CF of Metals		
	Square SP-4{90^4,180^2}	Pyramid SPY-5{90^6,105^2,165^2}	Octahedron OC-6{90^{12},180^3}
CHO_2^- L1 {S11d}	AB^2_2	$AB^2_2M^1$	$AB^2_2M^1_2$
A-2{135}	**sql** (30.8) **dia** (26.6) 2,4C4 (13.0)	2,4M6-3 (69.0) **sql** (7.6) 2,4C4 (6.8)	**sql** (58.6) 2,4C4 (13.6) **hcb** (7.2)
A-2{150}	**sql** (34.0) 2,4C4 (14.8) **dia** (14.2)	2,4M6-3 (63.8) **sql** (16.8) 2,4C4 (5.8)	**sql** (64.4) 2,4C4 (13.2) 2,4M6-3 (6.2)
$C_2H_3O_2^-$ L10 {S11e}	AB^2_2	$AB^2_2M^1$	$AB^2_2M^1_2$
A-2{45}	2,4C4 (34.2) **sql** (21.6) 2,4M6-3 (7.8)	2,4M6-3 (86.4) 2,4C4 (5.8) **sql** (3.2)	**sql** (47.9) 2,4M6-3 (16.0) 2,4C4 (13.1)
A-2{60}	2,4C4 (46.6) **sql** (24.6) 2,4M12-2 (3.8)	2,4M6-3 (56.4) 2,4C4 (21.4) 2,4M12-2 (13.0)	2,4C4 (37.8) **sql** (34.6) **hcb** (6.8)
$C_2O_4^{-2}$ L74 {S11j}	$AB^2M^1_2$	$AB^2M^1_3$	$AB^2M^1_4$
L-2 or A-2{180}	2C1 (76.8) 2M4-1 (13.8) (4,4)(0,2) (2.0)	2C1 (80.4) 2M4-1 (12.6) **bex** (1.8)	2C1 (98.6) 2M4-1 (0.6) 1,2,4M9-2 (0.2)
	AK^{22}	$AK^{22}M^1$	$AK^{22}M^1_2$
RAP-4{75^2,105^2,180^2}	**pts** (34.4) **cds** (27.6) **sql** (21.6)	**sra** (50.1) **sql** (33.7) 4,4L1 (7.6)	**sql** (50.4) 4,4,4T12 (12.0) **sra** (10.0)
$C_8H_4O_4^{-2}$ L81 {S11n}	AK^4	AK^4M^1	$AK^4M^1_2$
RAP-4{30^2,150^2,180^2}	**pts** (51.4) 4,4,4,6T40 (11.2) **sql** (10.8)	**pts** (52.4) 4,4L1 (35.6) (4,4)(2,2) (5.6)	**sql** (58.2) **dia** (10.6) **cds** (9.0)
RAP-4{45^2,135^2,180^2}	**pts** (64.4) 4,4,4,6T40 (10.8) **dia** (8.2)	**pts** (43.6) 4,4L1 (36.6) **sql** (9.1)	**sql** (55.8) **dia** (10.0) **cds** (8.4)
RAP-4{30^2,135,150,180^2}	**pts** (44.2) **sra** (16.2) 4,4,4,6T40 (11.0)	4,4L1 (36.4) **sql** (34.0) **pts** (20.8)	**sql** (56.2) **cds** (11.8) **dia** (10.4)