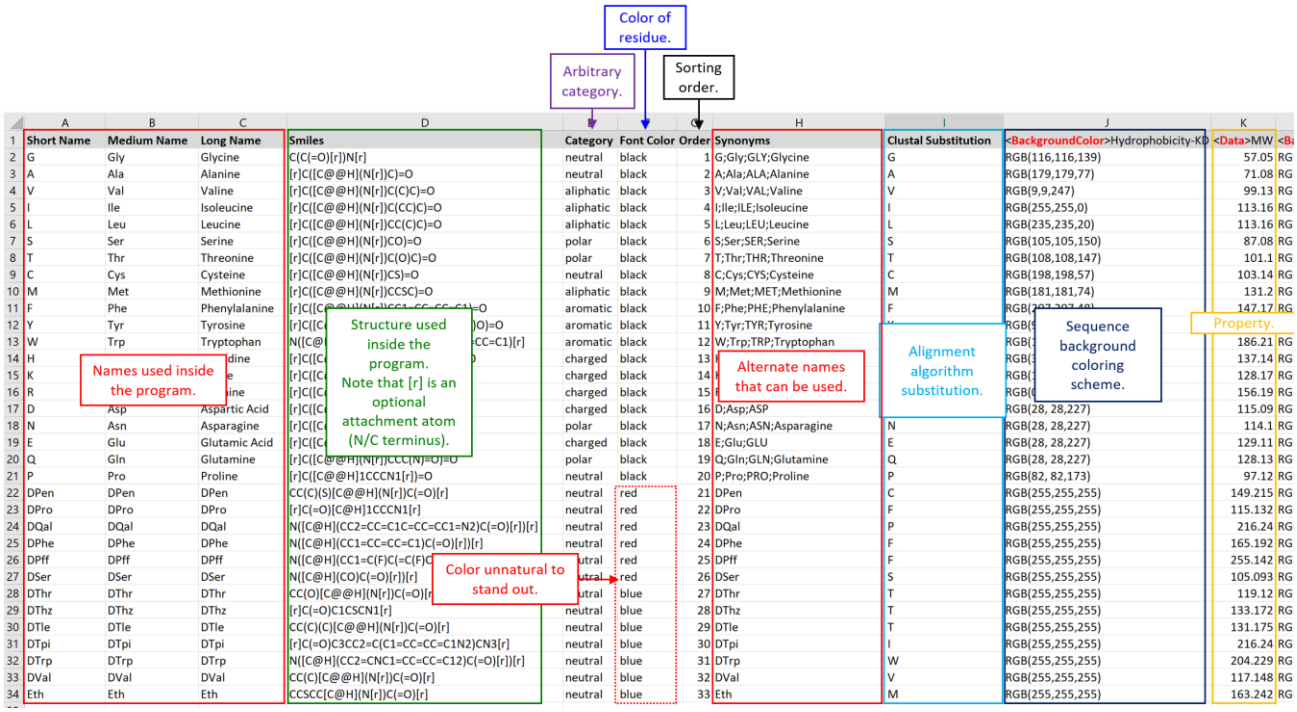
Notes on the Residue Table

SARvision|Biologics uses three files to create a data set. These include sequences and data (project file), residues (monomers used to build the biopolymer) and modifiers (transformations that are possible for resides in the residue table, such as methylation, pegylation, l to d). The two last ones, are control files that are modified occasionally as new residue types become part of the projects. Arguably, you can simply use the Residue Table and minimize the use of the Modifiers Table. Therefore, the Residue Table is at the heart of all the capabilities in SARvision|Biologics and deserves some further comments.

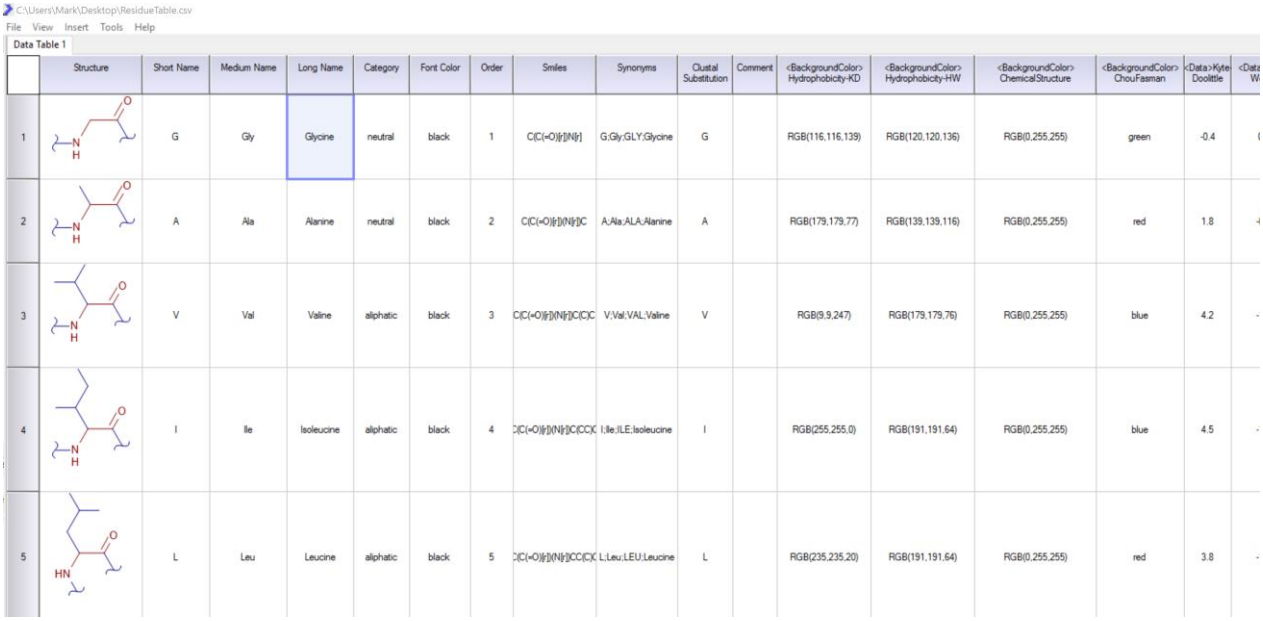
To find these control files: ***SARvision***: ***main menu: tools: open resource folder…*** These notes relate to the Monomer File that is

1. The Residue (Monomer) table example is ResidueTable.csv which is a comma delimited ascii file. It contains names, colors, structures (smiles) for each building block. This example is the 20 natural building blocks plus a number of unnatural amino acids.
   There are a number of important columns:
   a. Three name columns that contain a short, medium and long name. These are used interchangeably inside ***SARvision|Biologics*** to optimize the look of tables and views. Note that these can all be the same or of any string length. Ideally, '**Short Name**' < '**Medium Name**' < '**Long Name**'.
   b. The **Smiles** string is a molecule encoding for the chemical structure of the monomer. The column is necessary, but the individual entries can be empty. Note that Smiles strings can be generated for molecules using most chemically aware programs.
   c. **Category** is an arbitrary field used to group monomers inside the program for convenience. The labels are completely arbitrary; however, 'polar', 'aliphatic', 'charged', 'neutral', 'aromatic', 'crosslinking'…. are convenient.
   d. The **Font Color** is the color of the font for this residue in the program. These are usually black with red, blue, green…. To designate unnatural or otherwise interesting residues. Note that RGB(##,##,##) can be used instead of the common color names.
   e. **Order** column tells the program how to sort these residues. These numbers are completely arbitrary and up to the user's description. If the user sorts an alignment column then this number is used for the sorting order for the column.
   f. The **Synonyms** column contains other names for this residue that may appear in the data. For example Glycine may appear as 'GLY', 'Gly', or 'G'. I most well curated databases this would be a single entry.
   g. The **Clustal Substitution** is the residue that should be used for alignments. Note that most alignment matrices (Blossum, PAM….) only work with natural amino acids. There is capabilities to work with alignment matrices with more than 20 entries. Contact us for details. Note that 'X' designates unknown and is handled by the algorithms.
   h. Note that there can be any number of <**Data**> and <**BackgroundColor**> columns. These describe monomers: physico-chemical property and color respectively. For example, these can be used to color residues in the alignment table by hydrophobicity; or to color all the aromatic residues one color and the hydrophilic residues another color (e.g. clustal coloring).



	A	B	C	D	E	F	G	H	I	J	K
1	Short Name	Medium Name	Long Name	Smiles	Category	Font Color	Order	Synonyms	Clustal Substitution	<BackgroundColor>Hydrophobicity-KD	<Data>MW
2	G	Gly	Glycine	C(C(=O)[r])N[r]	neutral	black	1	G;Gly;GLY;Glycine	G	RGB(116,116,139)	57.05
3	A	Ala	Alanine	[r]C([C@@H](N[r])C)=O	neutral	black	2	A;Ala;ALA;Alanine	A	RGB(179,179,77)	71.08
4	V	Val	Valine	[r]C([C@@H](N[r])C(C)C)=O	aliphatic	black	3	V;Val;VAL;Valine	V	RGB(9,9,247)	99.13
5	I	Ile	Isoleucine	[r]C([C@@H](N[r])C(CC)C)=O	aliphatic	black	4	I;Ile;ILE;Isoleucine	I	RGB(255,255,0)	113.16
6	L	Leu	Leucine	[r]C([C@@H](N[r])CC(C)C)=O	aliphatic	black	5	L;Leu;LEU;Leucine	L	RGB(235,235,20)	113.16
7	S	Ser	Serine	[r]C([C@@H](N[r])CO)=O	polar	black	6	S;Ser;SER;Serine	S	RGB(105,105,150)	87.08
8	T	Thr	Threonine	[r]C([C@@H](N[r])C(O)C)=O	polar	black	7	T;Thr;THR;Threonine	T	RGB(108,108,147)	101.1
9	C	Cys	Cysteine	[r]C([C@@H](N[r])CS)=O	neutral	black	8	C;Cys;CYS;Cysteine	C	RGB(198,198,57)	103.14
10	M	Met	Methionine	[r]C([C@@H](N[r])CCSC)=O	aliphatic	black	9	M;Met;MET;Methionine	M	RGB(181,181,74)	131.2
11	F	Phe	Phenylalanine	[r]C([C@@H](N[r])CC1=CC=CC=C1)=O	aromatic	black	10	F;Phe;PHE;Phenylalanine	F	RGB(207,207,48)	147.17
12	Y	Tyr	Tyrosine	[r]C([C@@H](N[r])CC1=CC=C(O)C=C1)=O	aromatic	black	11	Y;Tyr;TYR;Tyrosine	Y	RGB(255,255,0)	163.17
13	W	Trp	Tryptophan	N([C@@H](N[r])CC1=CNC2=CC=CC=C12)[r]	aromatic	black	12	W;Trp;TRP;Tryptophan	W	RGB(180,90,180)	186.21
14	H			[r]C([C@@H](N[r])CC1=CN=CN1)=O	charged	black	13	H;His;HIS;Histidine	H	RGB(20,147,120)	137.14
15	K			[r]C([C@@H](N[r])CCCCN)=O	charged	black	14	K;Lys;LYS;Lysine	K	RGB(20,90,255)	128.17
16	R			[r]C([C@@H](N[r])CCCNC(N)=N)=O	charged	black	15	R;Arg;ARG;Arginine	R	RGB(20,90,255)	156.19
17	D	Asp	Aspartic Acid	[r]C([C@@H](N[r])CC(O)=O)=O	charged	black	16	D;Asp;ASP		RGB(28, 28,227)	115.09
18	N	Asn	Asparagine	[r]C([C@@H](N[r])CC(N)=O)=O	polar	black	17	N;Asn;ASN;Asparagine	N	RGB(28, 28,227)	114.1
19	E	Glu	Glutamic Acid	[r]C([C@@H](N[r])CCC(O)=O)=O	charged	black	18	E;Glu;GLU	E	RGB(28, 28,227)	129.11
20	Q	Gln	Glutamine	[r]C([C@@H](N[r])CCC(N)=O)=O	polar	black	19	Q;Gln;GLN;Glutamine	Q	RGB(28, 28,227)	128.13
21	P	Pro	Proline	[r]C([C@@H]1CCCN1[r])=O	neutral	black	20	P;Pro;PRO;Proline	P	RGB(82, 82,173)	97.12
22	DPen	DPen	DPen	CC(C)(S)[C@@H](N[r])C(=O)[r]	neutral	red	21	DPen	C	RGB(255,255,255)	149.215
23	DPro	DPro	DPro	[r]C(=O)[C@H]1CCCN1[r]	neutral	red	22	DPro	F	RGB(255,255,255)	115.132
24	DQal	DQal	DQal	N([C@H](CC2=CC=C1C=CC=CC1=N2)C(=O)[r])[r]	neutral	red	23	DQal	P	RGB(255,255,255)	216.24
25	DPhe	DPhe	DPhe	N([C@H](CC1=CC=CC=C1)C(=O)[r])[r]	neutral	red	24	DPhe	F	RGB(255,255,255)	165.192
26	DPff	DPff	DPff	N([C@H](CC1=C(F)C(=C(F)C	utral	red	25	DPff	F	RGB(255,255,255)	255.142
27	DSer	DSer	DSer	N([C@H](CO)C(=O)[r])[r]	utral	red	26	DSer	S	RGB(255,255,255)	105.093
28	DThr	DThr	DThr	CC(O)[C@@H](N[r])C(=O)[r	utral	blue	27	DThr	T	RGB(255,255,255)	119.12
29	DThz	DThz	DThz	[r]C(=O)C1CSCN1[r]	neutral	blue	28	DThz	T	RGB(255,255,255)	133.172
30	DTle	DTle	DTle	CC(C)(C)[C@@H](N[r])C(=O)[r]	neutral	blue	29	DTle	T	RGB(255,255,255)	131.175
31	DTpi	DTpi	DTpi	[r]C(=O)C3CC2=C(C1=CC=CC=C1N2)CN3[r]	neutral	blue	30	DTpi	I	RGB(255,255,255)	216.24
32	DTrp	DTrp	DTrp	N([C@H](CC2=CNC1=CC=CC=C12)C(=O)[r])[r]	neutral	blue	31	DTrp	W	RGB(255,255,255)	204.229
33	DVal	DVal	DVal	CC(C)[C@@H](N[r])C(=O)[r]	neutral	blue	32	DVal	V	RGB(255,255,255)	117.148
34	Eth	Eth	Eth	CCSCC[C@H](N[r])C(=O)[r]	neutral	blue	33	Eth	M	RGB(255,255,255)	163.242

2. These can be generated in a chemistry aware program such as SARvision-SM. It can be easily edited in excel and/or exported out of a database. Contact us if you want to directly integrate an oracle data table directly into SARvision-Biologics, such that the residue table is easily kept up to date with new monomer entries.



	Structure	Short Name	Medium Name	Long Name	Category	Font Color	Order	Smiles	Synonyms	Clustal Substitution	Comment	<BackgroundColor> Hydrophobicity-KD	<BackgroundColor> Hydrophobicity-HW	<BackgroundColor> ChemicalStructure	<BackgroundColor> ChouFasman	<Data>Kyte Doolittle	<Data> W
1		G	Gly	Glycine	neutral	black	1	C(C(=O)[r])[N[r]	G;Gly;GLY;Glycine	G		RGB(116,116,139)	RGB(120,120,136)	RGB(0,255,255)	green	-0.4	
2		A	Ala	Alanine	neutral	black	2	C([C(=O)[r])[N[r])C	A;Ala;ALA;Alanine	A		RGB(179,179,77)	RGB(139,139,116)	RGB(0,255,255)	red	1.8	
3		V	Val	Valine	aliphatic	black	3	C([C(=O)[r])[N[r])C(C)C	V;Val;VAL;Valine	V		RGB(9,9,247)	RGB(179,179,76)	RGB(0,255,255)	blue	4.2	
4		I	Ile	Isoleucine	aliphatic	black	4	C([C(=O)[r])[N[r])C(CC)C	I;Ile;ILE;Isoleucine	I		RGB(255,255,0)	RGB(191,191,64)	RGB(0,255,255)	blue	4.5	
5		L	Leu	Leucine	aliphatic	black	5	C([C(=O)[r])[N[r])CC(C)C	L;Leu;LEU;Leucine	L		RGB(235,235,20)	RGB(191,191,64)	RGB(0,255,255)	red	3.8	

3. The Modifier table is more simple. The entries can be pseudo-molecules ('Peg-30') or actual molecules (acetyl or 

). Note that there can be any number of <Data> columns. These entries are largely annotative in nature for visual purposes. These are not used for model building or mutation cliffs.

See ModifierTable.csv

A	B	C	D	E	F	G
Font Color	Name	Sort Order	Synonyms	Comment	<Data>MW	Smiles
red	acetyl	2	Ac;acetyl;Acetyl;		43.04	[r]C(C)=O
red	amide	3	amide;Amide		44.03	
red	lauroyl	4	Lauroyl;lauroyl		199.31	
red	myristoyl	5	Myristoyl;myristoyl		227.36	
red	palmitoyl	6	Pal;PAL;Palmitoyl;palmitoyl		255.42	
red	N-methyl	7	N-Me;N(Me);N(CH3);NNe;NCH3;N-methyl;N-Methyl		29.04	
black	N15	8	N15		1	
black	C14	9	C14		2	
black	C13	10	C13		1	
red	Peg-38atom	22	Peg-38 Atom;Peg-38a;Peg-38A;Peg-17atom;Peg-17Atom	MW=(12*2+1	5008	
red	Peg-17atom	23	Peg-17 Atom;Peg-17a;Peg-17A;Peg-17Atom;Peg-17atom	MW=(3*2+1)	1408	

Names | Possible names or aliases | Data | Optional Structure

4. Sequences can be read in many sequence formats and are usually in csv format with data. This example had an ID column, two possible sequence formats to read in, and 3 data columns. See GLP-1_sequences.csv.

ID	Sequence	Sequences(HELMS)	IC50	EC50	REF
GLP1-7137	HAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG	PEPTIDE1{H.A.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.R.G}	0.31	0.9	EJMC39:473(2004)
8	H[d-A]EGTFTSDVSSYLEGQAAKEFIAWLVKGG	PEPTIDE1{H.dA.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.G}	0.33	1.4	EJMC39:473(2004)
1	HAEGTFTSDVSSYLEGQAAKEFIAWLVKGG	PEPTIDE1{H.A.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.G}	0.18	0.5	EJMC39:473(2004)
2	FAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG	PEPTIDE1{F.A.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.R.G}	0.32	0.9	EJMC39:473(2004)
3	FAEGTFTSDVSSYLEGQAAKEFIAWLVKGG	PEPTIDE1{F.A.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.G}	1.6	7	EJMC39:473(2004)
4	WAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG	PEPTIDE1{W.A.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.R.G}	3.3	127	EJMC39:473(2004)
5	WAEGTFTSDVSSYLEGQAAKEFIAWLVKGG	PEPTIDE1{W.A.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.G}	4.6	152	EJMC39:473(2004)
6	YAEGTFTSDVSSYLEGQAAKEFIAWLVKGRG	PEPTIDE1{Y.A.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.R.G}	2.7	5.4	EJMC39:473(2004)
7	H[d-A]EGTFTSDVSSYLEGQAAKEFIAWLVKGRG	PEPTIDE1{H.dA.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.R.G}	0.15	0.8	EJMC39:473(2004)
9	HSEGTFTSDVSSYLEGQAAKEFIAWLVKGRG	PEPTIDE1{H.S.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.R.G}	2.9	15	EJMC39:473(2004)
10	HSEGTFTSDVSSYLEGQAAKEFIAWLVKGG	PEPTIDE1{H.S.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.G}	3.8	17	EJMC39:473(2004)
11	HVEGTFTSDVSSYLEGQAAKEFIAWLVKGRG	PEPTIDE1{H.V.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.R.G}	0.47	2.5	EJMC39:473(2004)
12	HVEGTFTSDVSSYLEGQAAKEFIAWLVKGG	PEPTIDE1{H.V.E.G.T.F.T.S.D.V.S.S.Y.L.E.G.Q.A.A.K.E.F.I.A.W.L.V.K.G.G}	1.4	3.5	EJMC39:473(2004)

5. Complex residues having more than a single letter designation should be added in brackets. Note that the single letters or the names in brackets should match the names in the Residue Table: Synonyms: column. Please contact us if you have a sequence format routinely used other than Fasta with brackets or HELMS. We can help with that.

ID	Name	Sequences	Sequences(HEI
Peptide-1	(Pyr1)-ape	[Pyr1]QRPRLSHKGPMPF	PEPTIDE1{[Pyr
1064		[pyrE]RPR[d-L]SHKGPMTY	PEPTIDE1{[pyr
1008		[pyrE]RPRLSKKG	PEPTIDE1{[pyr
1001		RPRLDHKGPM	PEPTIDE1{R.P.I
1002		RPRLSKKGPM	PEPTIDE1{R.P.I
1003		RPKLSHKGPM	PEPTIDE1{R.P.I
1004		RPR[d-L]SHKGPM	PEPTIDE1{R.P.I