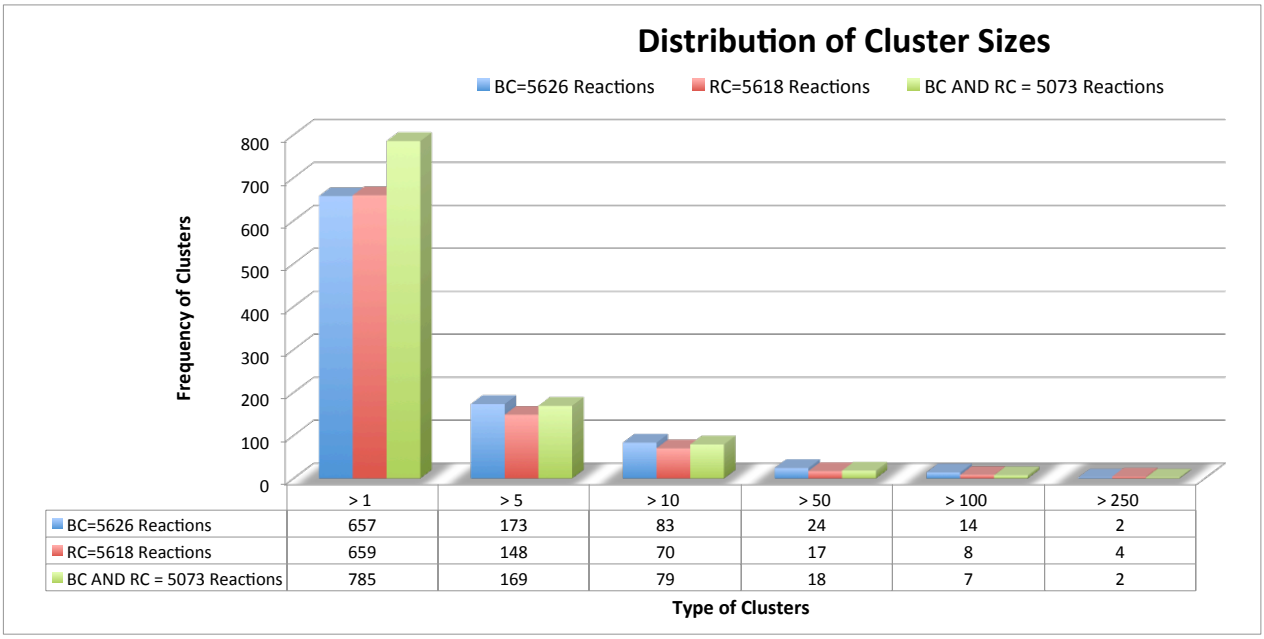
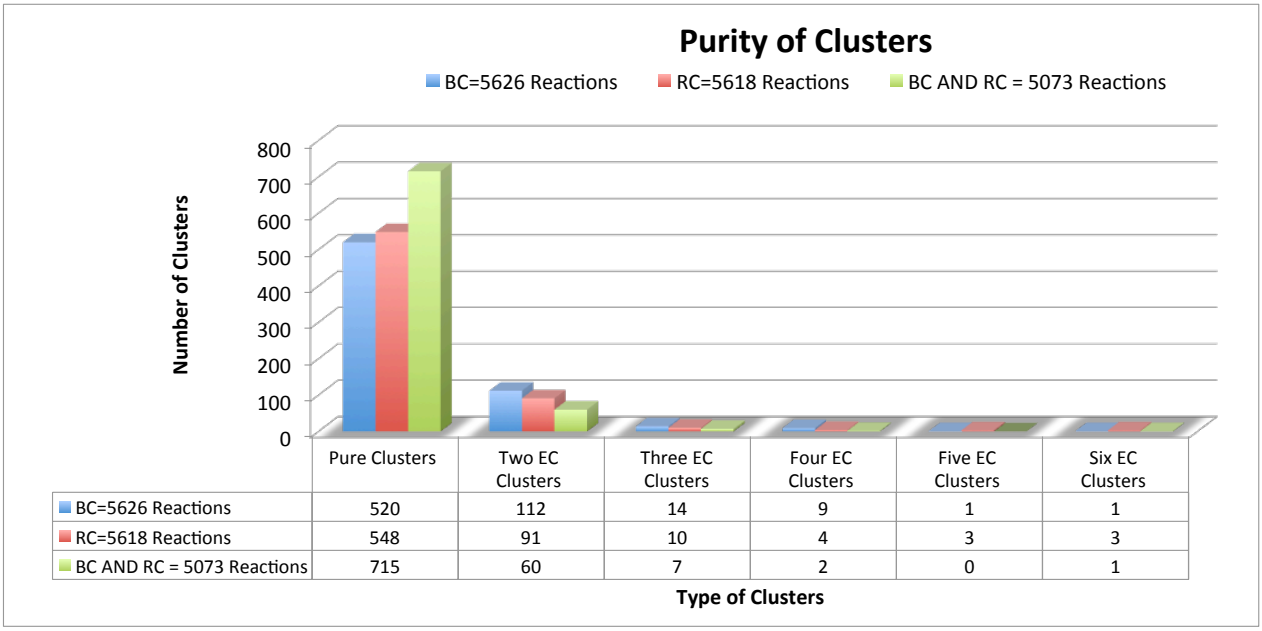
Supplementary Figure 1: Distribution of Bond Change and Reaction Centre Similarity clusters with p<0.01.



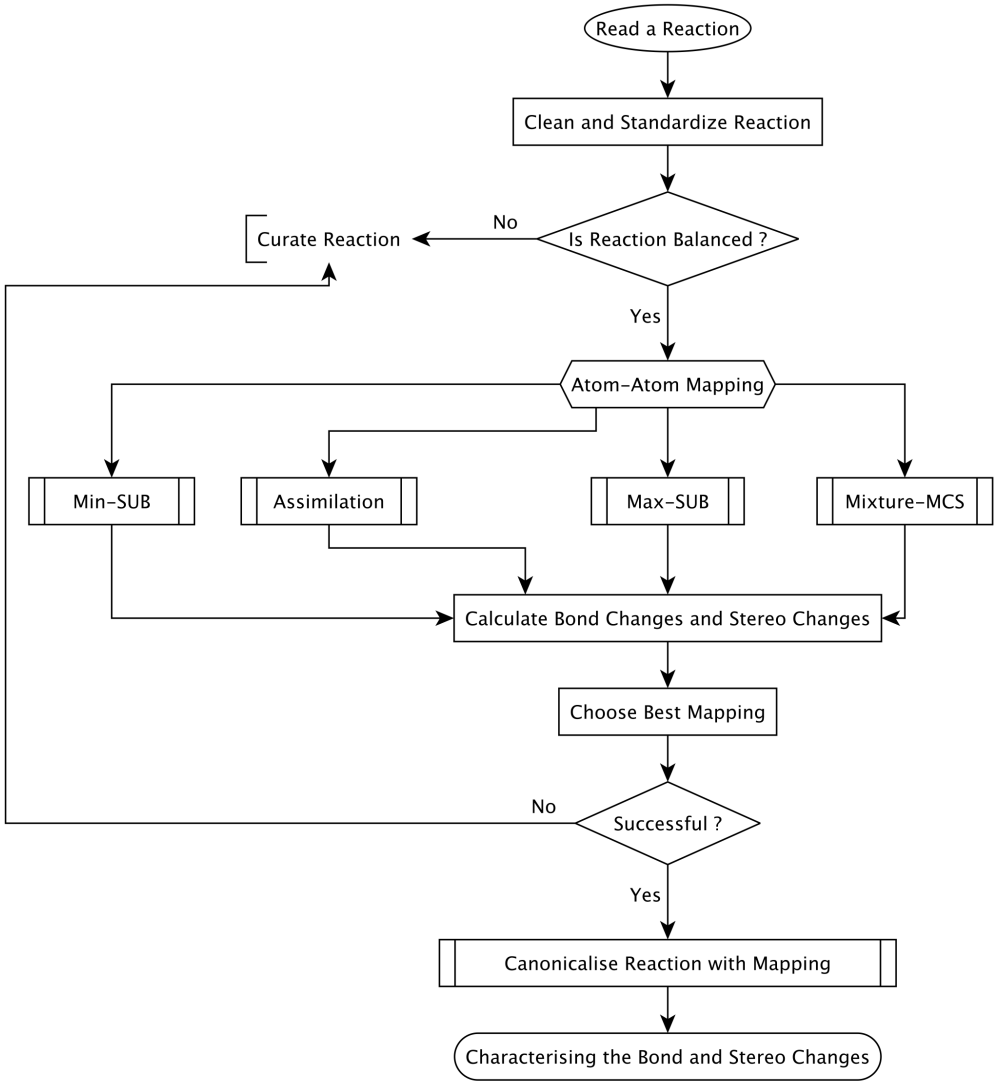
	> 1	> 5	> 10	> 50	> 100	> 250
BC=5626 Reactions	657	173	83	24	14	2
RC=5618 Reactions	659	148	70	17	8	4
BC AND RC = 5073 Reactions	785	169	79	18	7	2

Supplementary Figure 2: Purity of Bond Change and Reaction Centre Similarity clusters with p<0.01.



	Pure Clusters	Two EC Clusters	Three EC Clusters	Four EC Clusters	Five EC Clusters	Six EC Clusters
BC=5626 Reactions	520	112	14	9	1	1
RC=5618 Reactions	548	91	10	4	3	3
BC AND RC = 5073 Reactions	715	60	7	2	0	1

Supplementary Figure 3: Description of EC-BLAST algorithm. Flowchart for AAM and bond change assignment in a balanced reaction.



Supplementary Table 1: Combination of Bond Change and Reaction Centre Similarity Network of Enzymes with p < 0.01.

Measure Similarity (p<0.01)	Bond Change Network	Reaction Centre Network	Bond Change and Reaction Centre
Total Nodes (Reactions)	5626	5618	5073
Total Edges	285095	204177	118014
Average Connectivity	101	72	46

Supplementary Table 2: Enzymes reported in the Phosphatidylinositol Phosphodiesterase (PPI) Domain Superfamily 3.20.20.190 (CATH version 3.5). The enzymes (EC) highlighted in yellow are balanced and at least one representative reaction is found in the EC-BLAST. EC highlighted in cyan are unbalanced hence they are not included in the analysis.

ID	Description	Reaction
2.4.1.187	N-acetylglucosaminyldiphosphoundecaprenol N-acetyl-beta-D-mannosaminyltransferase.	UDP-N-acetyl-D-mannosamine + N-acetyl-D-glucosaminyldiphosphoundecaprenol = UDP + N-acetyl-beta-D-mannosaminyl-1,4-N-acetyl-D-glucosaminyldiphosphoundecaprenol.
2.7.7.48	RNA-directed RNA polymerase.	Nucleoside triphosphate + RNA(n) = diphosphate + RNA(n+1).
3.1.1.5	Lysophospholipase.	2-lysophosphatidylcholine + H(2)O = glycerophosphocholine + a carboxylate.

3.1.22.1	Deoxyribonuclease II.	Endonucleolytic cleavage to nucleoside 3'-phosphates and 3'-phosphooligonucleotide end-products.
3.1.3.1	Alkaline phosphatase.	A phosphate monoester + H(2)O = an alcohol + phosphate.
3.1.4.11	Phosphoinositide phospholipase C.	1-phosphatidyl-1D-myo-inositol 4,5-bisphosphate + H(2)O = 1D-myo-inositol 1,4,5-trisphosphate + diacylglycerol.
3.1.4.13	Serine-ethanolaminephosphate phosphodiesterase.	Serine phosphoethanolamine + H(2)O = serine + ethanolamine phosphate.
3.1.4.3	Phospholipase C.	A phosphatidylcholine + H(2)O = 1,2-diacylglycerol + choline phosphate.
3.1.4.4	Phospholipase D.	A phosphatidylcholine + H(2)O = choline + a phosphatidate.
3.1.4.41	Sphingomyelin phosphodiesterase D.	Sphingomyelin + H(2)O = ceramide phosphate + choline.
3.1.4.43	Glycerophosphoinositol inositolphosphodiesterase.	1-(sn-glycero-3-phospho)-1D-myo-inositol + H(2)O = glycerol + 1D-myo- inositol 1-phosphate.
3.1.4.44	Glycerophosphoinositol glycerophosphodiesterase.	1-(sn-glycero-3-phospho)-1D-myo-inositol + H(2)O = myo-inositol + sn-glycerol 3-phosphate.
3.1.4.46	Glycerophosphodiester phosphodiesterase.	A glycerophosphodiester + H(2)O = an alcohol + sn-glycerol 3-phosphate.
3.2.1.14	Chitinase.	Random hydrolysis of N-acetyl-beta-D-glucosaminide (1->4)-beta-linkages in chitin and chitodextrins.
3.2.1.73	Licheninase.	Hydrolysis of (1->4)-beta-D-

		glucosidic linkages in beta-D-glucans containing (1->3)- and (1->4)-bonds.
3.4.11.4	Tripeptide aminopeptidase.	Release of the N-terminal residue from a tripeptide.
4.6.1.13	Phosphatidylinositol diacylglycerol-lyase.	1-phosphatidyl-1D-myo-inositol = 1D-myo-inositol 1,2-cyclic phosphate + 1,2-diacyl-sn-glycerol.
4.6.1.14	Glycosylphosphatidylinositol diacylglycerol-lyase.	6-(alpha-D-glucosaminyl)-1-phosphatidyl-1D-myo-inositol = 6-(alpha-D- glucosaminyl)-1D-myo-inositol 1,2-cyclic phosphate + 1,2-diacyl- sn-glycerol.

Supplementary Table 3: Using EC-BLAST to find reactions of related sequences in the Phosphatidylinositol Phosphodiesterase (PPI) Domain Superfamily. Starting with the most prevalent EC number 3.1.4.46 (which shares its reactions with EC 3.1.4.2) in the PPI superfamily as the search term in EC-BLAST, the resulting EC numbers found in the EC-BLAST hit list are shown (only the top 100 results are shown). This search was performed on the Reaction centre (RC) and Structure Similarity (SS) metric. Some of the reactions have R-group present in them. "*" Indicates hits using both reaction centre and structure similarity.

IUBMB EC (CATH 3.20.20.190)	**RC Ranking**	**IUBMB EC (CATH 3.20.20.190)**	**SC Ranking**
3.1.4.46	Query	**3.1.4.46**	Query
3.1.4.41	2	**3.1.1.5**	5
3.1.4.4	4*	**3.1.4.44**	17

3.1.4.3	5*	**3.1.4.13**	21*
3.1.4.13	7*	**3.1.4.3**	32*
4.6.1.13	12	**3.1.4.43**	49
3.1.3.1	33*	**3.1.4.4**	57*
		3.1.3.1	84*

SUPPLEMENTARY RESULTS

Reaction Similarity Network

The reactions present in Fig. 3c case i, R07059 (EC 1.1.1.271, GDP-L-fucose synthase), R08597 (EC 5.1.3.-, dTDP-4-oxo-2,6-dideoxy-D-glucose 3,5-epimerase) and R06514 (EC 5.1.3.13, dTDP-4-dehydrorhamnose 3,5-epimerase), undergo a shared mechanism of epimerisation of a nucleotide-mannose by enzymatic abstraction of the C3 and C5 protons. In case ii, reactions R06974 (EC 2.1.2.-, glycinamide ribonucleotide transformylase), R06975 (EC 6.3.4.-, 5-formaminoimidazole-4-carboxamide-1-beta-D-ribofuranosyl 5'-monophosphate synthetase) and R02238 (EC 6.3.4.17, formate—dihydrofolate ligase) belonging to two different primary classes share the same mechanism of amide formation using ATP. In many such mixed clusters only one of the reactions will belong to a different class and such outlier reaction may be errors, or more likely reactions with attributes of both the primary classes represented (for example a 5.3 reaction which is an intramolecular oxidoreductase isomerase). This highlights the challenge of assigning all reactions, many of which are complex, into the hierarchical EC classification scheme.

Finding Reactions of Related Sequences in the Phosphatidylinositol Phosphodiesterase (PPI) Domain Superfamily

It is well known that many enzymes duplicate and then evolve to perform many different enzyme functions [1-5], for example Nguyen et *al*. [6] reported mechanistic evidence on how a uronate isomerase activity might have evolved from a hydrolase activity within the amidohydrolase (AHS) superfamily by modification of the enzyme active site. The reaction similarity between two closely related enzymes sequences can be very low in some cases [7]. At the other extreme two unrelated sequences can perform the same overall reaction (e.g. the chloroperoxidases [8]). Thus, it is not always possible to infer reaction similarity from sequence similarity. However, it is reasonable

to assume that, in general, enzyme function will evolve slowly [9] and it is therefore likely that the reactions of closely related enzymes will show some reaction similarity by retaining the substrate/product specificity or the mechanism of the reactions they catalyse.

We catalogued 8,823 sequences that contain the PPI domain (CATH: 3.20.20.190) as reported by CATH-Gene3D [10,11] version 3.5. The resulting enzymes perform 18 different enzyme functions of which 12 can be included in the EC-BLAST database (Supplementary Table 2). Five of these are not included in EC-BLAST as the reactions are unbalanced (EC 2.7.7.48, EC 3.1.22.1, EC 3.2.1.14, EC 3.2.1.73, and EC 3.4.11.4). The most prevalent EC number within this superfamily EC 3.1.4.46 (glycerophosphodiester phosphodiesterase) was associated with 1204 (77%) curated sequences. To find the most closely related reactions from a chemical perspective, this reaction (KEGG Reaction R01030) was used as the search term in EC-BLAST (**Supplementary Table 3**). The goal is to explore if this enzyme superfamily has evolved over time to include family members, which perform any of these most closely related reactions. 18 IUBMB-EC members of the PPI family as defined by the CATH domain superfamily (CATH 3.20.20.190) have KEGG reactions assigned to it. Five enzymes have unbalanced / incomplete reaction(s), hence they are not included in the analysis (EC 2.7.7.48, EC 3.1.22.1, EC 3.2.1.14, EC 3.2.1.73, and EC 3.4.11.4).

Using the reaction centre metric 6 of the possible other 12 reactions performed by this superfamily are found in the top fifty EC-BLAST results list. Of these 4 were in the top 7, being very similar reactions in the same EC sub-subclass. The fifth, which is a lyase rather than an oxidoreductase, comes in at rank 12. Three reactions performed by this superfamily were not identified in these searches (EC 2.4.1.187, EC 3.1.4.11, EC 4.6.1.14). Although we were able to match EC 4.6.1.13 with the query due to the presence of water performing hydrolysis in the reaction, this reaction centre is absent in EC 4.6.1.14, hence this reaction does not appear in the top 100 list of EC-BLAST. Likewise EC 3.1.4.11 doesn't appear in top 100 due to the presence of R-group in the reaction centre (data artefact). If this reaction is used as a search

term then EC 3.1.4.46 appears as one of the top 50 hits. All the enzymes in this family share O-H & O-P bond changes except for enzyme (EC 2.4.1.187), which catalyses changes in O-C & O-H bonds. Just 9 sequences in UniProt [12] are annotated with this function which therefore appears to be rather rare.

SUPPLEMENTARY NOTES 1: Algorithm

Describing the Chemical Reaction

a. Let a chemical reaction with 'n' reactants and 'm' products can be defined by

Set of reactants $S = \{s_n | n \in \mathbb{N}_1\}$ and
Set of products $P = \{p_m | m \in \mathbb{N}_1\}$.

b. Each reactant (S) and product (P) can further be broken down into set of atoms (a) and bonds. In a balanced reaction the number of atoms on the reactant side is equivalent to the number of atoms on the product side.

Let A_S and A_P represent set of atoms in reactants (S) and products (P).

$A_S = \{a_i | i \in \mathbb{N}_1, a \in A_S\}$

$A_P = \{a_j | j \in \mathbb{N}_1, a \in A_P\}$

A reaction can be represented as an order isomorphism (π) set with a $bijective$ function $h: A_S \rightarrow A_p$, where $h(a_i) \leq_{A_S} h(a_j)$ $if\ and\ only\ if\ a_i \leq_{A_P} a_j$.

Let B_S and B_P represents set of bonds in reactants (S) and products (P).

$B_S = \{b_k^l | k, l \in \mathbb{N}_0, k < l\ and\ b \in B_S\}$

$B_P = \{b_p^q | p, q \in \mathbb{N}_0, p < q\ and\ b \in B_P\}$

$In\ a\ balanced\ reaction$:

$number\ of\ atoms\ in\ the\ reactants = number\ of\ atoms\ in\ the\ products.$

Hence in a balanced reaction ($where\ i = j\ and\ |A_S| = |A_P|,$) the one to one mapping of the atoms between reactant and product atoms are called $Atom - Atom\ Mapping\ (AAM)$.

This can be defined by using $bijective$ function $f: A_S \rightarrow A_P \Leftrightarrow$ it satisfies the condition for every $a_j \in A_P$ there is a unique $a_i \in A_S$ with $a_j = f\ (a_i)$.

Graph Matching

a. A molecular graph using atoms and bonds can be represented as a labelled graph G.

$Let\ G = \{V, E, l_a, l_b\}$
$where\ V: atoms\ of\ the\ molecule,$
$E: bonds\ of\ the\ molecule,$
$l_a and\ l_b\ represents\ atom\ types\ and\ bond\ types\ respectively.$

$Let\ G_s - reactant\ graph\ and\ G_p - product\ graph,$

b. Represent a graph of reactant (S) and product (P) molecules respectively. The $Maximum\ Common\ Subgraph\ (MCS)$ between two molecules using graph theory can be defined as

Let G_H represent MCS between $G_s\ and\ G_p\ \ if\ and\ only\ if\ G_H\ is$ a subgraph of $G_s\ and\ G_p\ and$ there is no graph which is a $subgraph$ of $G_s\ and\ G_p,$ including strictly more vertices than G_H.

Overall Atom-Atom Mapping Procedure

Generic, pre-mapping steps
$Let\ (m = 3\ reactants)\ and\ (n = 2\ products)$:

1. $Given\ a\ reaction\ s_1 + s_2 + s_3 \leftrightarrow p_1 + p_2$
2. $Canonicalise\ and\ index\ the\ reactants\ and\ products\ based\ on\ the\ rules.\ See\ supplementary\ section\ 6\ and\ 7\ for\ canonicalisation\ rules.$
3. $Generate\ isomorphism\ (\pi) similarity\ matrix\ between\ reactant\ and\ product\ matrix.$
4. $Compare\ all\ molecules\ on\ each\ side\ of\ the\ equation,\ (s_1 \rightarrow p_1;\ s_1 \rightarrow p_2;\ s_2 \rightarrow p_2;\ s_2 \rightarrow p_3;\ s_3 \rightarrow p_1;\ s_3 \rightarrow p_1)$
5. $for\ each$ comparison identify the MCS using SMSD.
6. $Calculate$ Tanimoto score $(T_s)\ for\ that\ match.$
7. $Populate\ the\ m * n\ matrix\ with\ the\ T_s\ similarity\ score.$
8. $Next\ step\ is\ dependent\ on\ the\ chosen\ model$
9. $if\ unmapped\ atoms\ exists\ then\ go\ to\ step\ 1\ else\ stop.$

Chemical Similarity Function:

The chemical similarity in the EC-BLAST is represented in two forms [13]:

a. Tanimoto score (T_s): The similarity between two binary vectors (x_i, y_i) of length $\mathcal{F}$ can be defined as

$$T_s(X_i, Y_i) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

Equation 1S: Tanimoto similarity

b. Weighted Jaccard Co-efficient (T_w)**:** The similarity between two weighted descriptor vector (v_{Z_i} and v_{Z_j}) of length $\mathcal{F}$ can be defined as

$$T_w(Z_i, Z_j) = \frac{\sum_{x=1}^{|\mathcal{F}|} \left(v_{Z_i} * v_{Z_j}\right)}{\sum_{x=1}^{|\mathcal{F}|} (v_{Z_i})^2 + \sum_{x=1}^{|\mathcal{F}|} \left(v_{Z_j}\right)^2 - \sum_{x=1}^{|\mathcal{F}|} \left(v_{Z_i} * v_{Z_j}\right)}$$

Supplementary Equation 2: Weighted Jaccard Similarity

Deadlock Resolver:

If two cells in the similarity matrix have the same score then

- The mapping, which generates the lesser number of fragments, if the mapped substructure is taken off the query molecules, is preferred.

- If there is still a deadlock or clash then the mapping, which produces minimum number of stereo changes and bond energy, is chosen.

Molecule Canonicalisation:

The concept of molecule canonicalisation is to uniquely represent a molecule's tautomeric forms. A molecule can be drawn in several ways and thus the atom order will change with each drawing. In order to maintain a consistency in the atom ordering we canonicalise the molecule using a modified form of molecular signatures [14]. This concept ranks the atom by its atom types (symbols and its hybridization states) and its connectivity (degree of the vertex).

The most standard way to canonicalise a molecule would be to convert it to InChI [13,15] format and back but presently InChI doesn't support pseudo atom types like "R"- groups.

Canonicalisation helps us to create a non-redundant representation of the molecules in our database.

Reaction Canonicalisation:

Like molecules, reactions too can be drawn in various ways and the order of the molecules can change accordingly. Thus the same reaction can produce

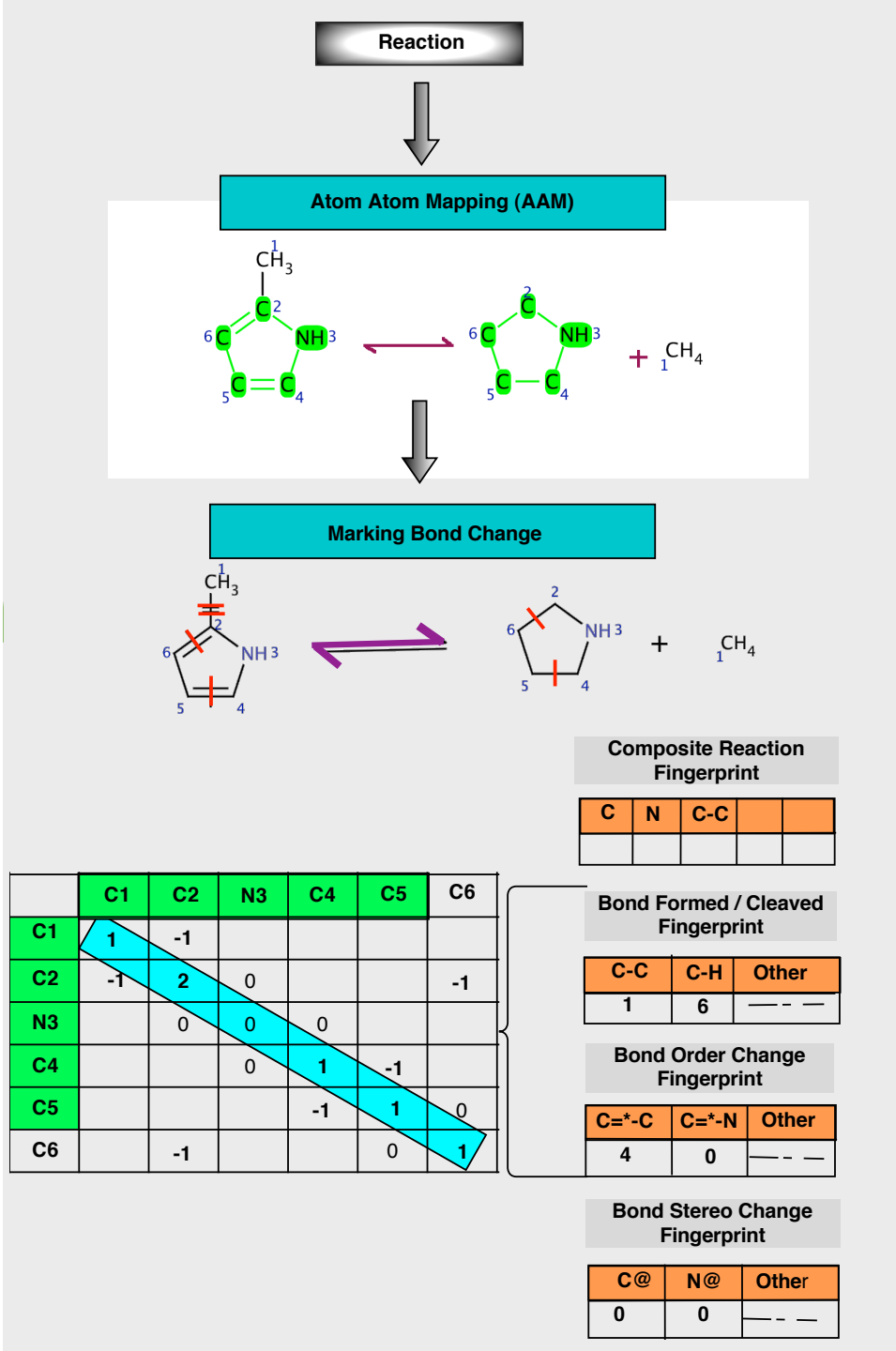
different atom-atom mapping (AAM) numbering order. This might results in confusion while referring back to the original atom. The goal of the reaction canonicalisation is to maintain the consistency in the AAM numbering such that successive runs of the same reactions can produce identical AAM.

Canonicalisation steps:

a. Rearrange the reactants and products in the order (atom count, labelling rank) of their size.
b. Canonicalise the molecules.
c. Perform the AAM mapping.
d. Re-number the AAM with respect to the arrangement of the atoms in the reactant molecules and their corresponding sub-graphs in the products.

Automated Extraction of the Bond Changes:

Figure A: Automated extraction of the bond changes from R-matrix (DU-Model) obtained from the AAM of the reactions.



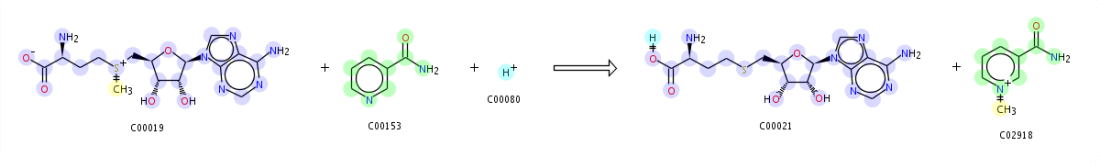
	C1	C2	N3	C4	C5	C6
C1	1	-1				
C2	-1	2	0			-1
N3		0	0	0		
C4			0	1	-1	
C5				-1	1	0
C6		-1			0	1

The R-matrix is transformed into bond change fingerprints (**Supplementary Fig.** A). The R-matrix is slightly modified and the diagonal can now store the number of hydrogens attached (protons changed) to an atom rather than the free valence electrons.

Mining Bond Changes:

The bond changes in a nicotinamide N-methyltransferase reaction (**Supplementary Fig. B**) are marked using the DU model after the AAM process. The annotated bond changes can be transformed as a bond change fingerprints i.e. bonds cleaved/formed {H-O (1), C-N (1), C-S (1)}, no bond order or stereo changes. We also store information about bonds being in the ring system, aliphatic or aromatic etc.

Figure B: An example reaction with marked bond changes and the mapped substrates and products from EC-BLAST. Substructure matches between the reactant and products molecules share identical colour code.

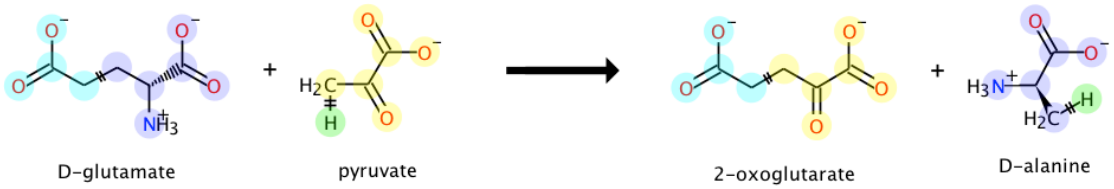


Mapping Challenges:

Some reactions do not follow the Principle of Minimum Chemical Distance (PMCD) [14,16] in terms of bond changes (**Supplementary Fig. C**). By comparing EC-BLAST substrate-product pair mappings with RPAIR we found 274 mismatches. In ~80% of the 274 failed cases, correct mapping was found by one of our algorithms, but was not chosen by the optimisation algorithm. In ~20% of the 274 failed cases KEGG solution was not found at all and in some cases the KEGG mapping was ambiguous. For example, using our mapping

algorithm, the KEGG reaction 2-oxoglutarate aminotransferase (R01148) can be mapped in two possible ways. The best solution, based on the lowest energy and minimal bond changes, is shown in mapping case 1 (**Supplementary Fig. C**). However information available from the literature and in the MACiE database (entry M0066) indicates that in fact mapping case 2 (**Supplementary Fig. C**) is correct. Any automated algorithm will not find such complex mappings, which contradict our assumption of minimal changes [16], unless additional information, especially the correct protonation states of the reactants and amino acids involved in the catalytic mechanism are included.

Figure C: Challenges in mapping KEGG reaction 2-oxoglutarate aminotransferase R01148. The best solution based on the lowest energy and minimal bond changes (mapping case 1) chosen by our ranking scheme is biochemically less favourable than mapping case 2, which has more bond changes, more fragments and a larger change in bond energy. Matching substructures share similar colour codes: yellow, cyan, light green, light pink. Hydrogens attached to carbons are only shown if they change their bonding.



CASE 1: Algorithm: Minimisation, Maximisation, Assimilation

Delta Energy: 692.0, Fragments: 2

Fingerprint Cleaved/Formed: C-C (2), C-H (2)

CASE 2: Algorithm: Mixture

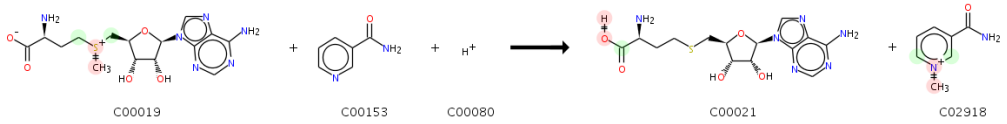
Delta Energy: 2208.0, Fragments: 4

Fingerprint Cleaved/Formed: C=0 (2), C-H (2), C-N (2)

Mining Reaction Centre Changes:

The reaction centres in the nicotinamide N-methyltransferase reaction are marked using the DU model after the AAM process (**Supplementary Fig. D**).

Figure D: The reaction centres are marked as red and the interacting atoms are marked as green.

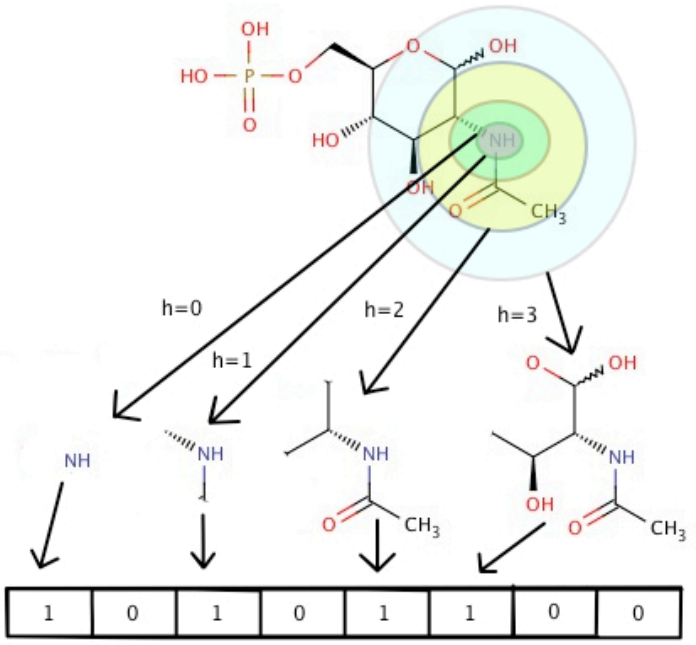


These reaction centres can be transformed into fingerprint patterns using the signature based circular fingerprints.

Circular Fingerprints:

The circular fingerprints (**Supplementary Fig. E**) can capture the atom environments around the atom of interest [14]. The size of the substructure can be manipulated by the diameter of the circle (h=height of the atom signature) around the central atom [14]. We have mixed the circular fingerprint with atom signatures to generate a canonicalized set of substructure patterns. These patterns can then be loaded as fingerprints of fixed size using a hashing algorithm.

Figure E: The circular fingerprint captures the neighbourhood information around the atom of interest. The size of the substructure can be controlled by the diameter/height of the circular search space.



Supplementary Note 2: R Script

```
library(ROCR)


my_P_Z_function <- function(scores)
{
	#Z-Score
	z<-scale(scores, center = TRUE, scale = TRUE)
	#p-value using extreme value distribution (EVD)
	p<-1-exp(-exp(-z*pi/sqrt(6)-0.5772157))
	return(list(r1 = z, r2 = p))
}


#Input Data Type Example
#  V1 V2    V3    V4    V5    V6
#1  2  1 1.000  true  true  true
#2  2  0 0.200 false false  true



my_assign_combinations<-function(bondchanges,reactioncenter, structure)
{
all.df<-data.frame(I = bondchanges$V1, J = bondchanges$V2)
all.df$ECX<-bondchanges$V4
all.df$ECX.X<-bondchanges$V5
all.df$ECX.X.X<-bondchanges$V6
all.df$BC<-bondchanges$V3
all.df$RC<-reactioncenter$V3
all.df$ST<-structure$V3


#assign Z-Scores
all.df$BC_Z<-scale(all.df$BC, center = TRUE, scale = TRUE)
all.df$RC_Z<-scale(all.df$RC, center = TRUE, scale = TRUE)
all.df$ST_Z<-scale(all.df$ST, center = TRUE, scale = TRUE)

#assign p-values
all.df$BC_P<-1-exp(-exp(-all.df$BC_Z*pi/sqrt(6)-0.5772157))
all.df$RC_P<-1-exp(-exp(-all.df$RC_Z*pi/sqrt(6)-0.5772157))
all.df$ST_P<-1-exp(-exp(-all.df$ST_Z*pi/sqrt(6)-0.5772157))

#assign combinations
all.df$BC_RC<-sqrt(exp(all.df$BC+all.df$RC)/2)
all.df$BC_ST<-sqrt(exp(all.df$BC+all.df$ST)/2)
all.df$RC_ST<-sqrt(exp(all.df$RC+all.df$ST)/2)

all.df$BC_RC_Z<-scale(all.df$BC_RC, center = TRUE, scale = TRUE)
all.df$BC_ST_Z<-scale(all.df$BC_ST, center = TRUE, scale = TRUE)
all.df$RC_ST_Z<-scale(all.df$RC_ST, center = TRUE, scale = TRUE)

all.df$BC_RC_P<-1-exp(-exp(-all.df$BC_RC_Z*pi/sqrt(6)-0.5772157))
all.df$BC_ST_P<-1-exp(-exp(-all.df$BC_ST_Z*pi/sqrt(6)-0.5772157))
all.df$RC_ST_P<-1-exp(-exp(-all.df$RC_ST_Z*pi/sqrt(6)-0.5772157))

return (all.df)
}
```

```
fileBC<-"data/RawDataBC.csv"
fileRC<-"data/RawDataRC.csv"
fileST<-"data/RawDataST.csv"

dataBC<-read.table(fileBC,sep=",", header=FALSE)
dataRC<-read.table(fileRC,sep=",", header=FALSE)
dataST<-read.table(fileST,sep=",", header=FALSE)

combine<-my_assign_combinations(dataBC,dataRC,dataST)


#########################################################
#PLOT GENERATION
#########################################################

#########################################################
#ROC curve Plot for EC Sub-SubClass
#########################################################

filepng<-"ROC_EC_SUB_SUB_CLASS.pdf"
#title( "EC-BLAST Scores", outer = TRUE )


pred_RC<-prediction(combine$RC, as.logical(combine$ECX) & as.logical(combine$ECX.X) & as.logical(combine$ECX.X.X))
pred_BC<-prediction(combine$BC, as.logical(combine$ECX) & as.logical(combine$ECX.X) & as.logical(combine$ECX.X.X))
pred_ST<-prediction(combine$ST, as.logical(combine$ECX) & as.logical(combine$ECX.X) & as.logical(combine$ECX.X.X))

plot(performance(pred_RC, measure = "tpr", x.measure = "fpr"), avg='threshold', spread.estimate='stddev', lwd=2.5, col="cornflowerblue",
main="ROC Curve for Tanimoto scores vs. EC Sub-SubClass Matches")
plot(performance(pred_BC, measure = "tpr", x.measure = "fpr"), avg='threshold', spread.estimate='stddev', lwd=2.5, col="DARKRED", add=TRUE)
plot(performance(pred_ST, measure = "tpr", x.measure = "fpr"), avg='threshold', spread.estimate='stddev', lwd=2.5, col="darkorange", add=TRUE)

legend("bottomright", c("Bond changes","Reaction center", "Structure similarity"), lty=c(1,1,1), lwd=c(2.5, 2.5, 2.5), col=c("DARKRED","cornflowerblue", "darkorange"))

quartz.save(filepng,type="pdf",device=dev.cur(),dpi=100)
dev.off()


#########################################################
#Accuracy Plot for EC Sub-Sub Class
#########################################################

filepng<-"Accuracy_EC_SUB_SUB_CLASS.pdf"

plot(performance(pred_RC, measure = "acc"), lwd=2.5, col="cornflowerblue",
main="Accuracy Curve for Predicting \nEC Sub-SubClass Matches using Jaccard", xlab="Jaccard cut-off score")
plot(performance(pred_BC, measure = "acc"), lwd=2.5, col="DARKRED", add=TRUE)
plot(performance(pred_ST, measure = "acc"), lwd=2.5, col="darkorange", add=TRUE)

legend("bottomright", c("Bond changes","Reaction center", "Structure similarity"), lty=c(1,1,1), lwd=c(2.5, 2.5, 2.5), col=c("DARKRED","cornflowerblue", "darkorange"))
```

```
quartz.save(filepng,type="pdf",device=dev.cur(),dpi=100)
dev.off()


############################################################
#CHI-SQ Plot for EC Sub-Sub Class
############################################################

filepng<-"CHI_SQ_EC_SUB_SUB_CLASS.pdf"

plot(performance(pred_RC, measure = "chisq"), lwd=2.5, col="DARKRED",
main="Chi square test statistic for Tanimoto scores vs. EC Sub-SubClass Matches")
plot(performance(pred_BC, measure = "chisq"), lwd=2.5, col="cornflowerblue", add=TRUE)
plot(performance(pred_ST, measure = "chisq"), lwd=2.5, col="darkorange", add=TRUE)

legend("bottomright", c("Bond changes","Reaction center", "Structure similarity"), lty=c(1,1,1),
lwd=c(2.5, 2.5, 2.5), col=c("DARKRED","cornflowerblue", "darkorange"))

quartz.save(filepng,type="pdf",device=dev.cur(),dpi=100)
dev.off()


############################################################
#Vioplot Plot for EC Sub-Sub Class
############################################################

library(vioplot)
filepng<-"Score_Density_EC_SUB_SUB_CLASS.pdf"


vioplot(combine$BC, combine$RC, combine$ST, names=c("Bond changes","Reaction center",
"Structure similarity"), col="GOLD")
title("Similarity score distribution", xlab="Similarity types", ylab="Jaccard score")
quartz.save(filepng,type="pdf",device=dev.cur(),dpi=100)
dev.off()



############################################################
#Combination p<-0.05
#Significance Data (p<0.05) or 1% chance of error, (p<0.05) or 5% chance of error
############################################################



combine_BC_0_05<-subset(combine, combine$BC_Z > 0. & combine$BC_P < 0.05)
combine_RC_0_05<-subset(combine, combine$RC_Z > 0. & combine$RC_P < 0.05)
combine_ST_0_05<-subset(combine, combine$ST_Z > 0. & combine$ST_P < 0.05)
combine_BC_RC_0_05<-subset(combine, combine$BC_Z > 0. & combine$BC_P < 0.05 &
combine$RC_Z > 0. & combine$RC_P < 0.05)
combine_BC_ST_0_05<-subset(combine, combine$BC_Z > 0. & combine$BC_P < 0.05 &
combine$ST_Z > 0. & combine$ST_P < 0.05)
combine_RC_ST_0_05<-subset(combine, combine$RC_Z > 0. & combine$RC_P < 0.05 &
combine$ST_Z > 0. & combine$ST_P < 0.05)
```

```
##########################################################
#Accuracy Plot for EC Sub-Sub Class
##########################################################

filepng<-"Combined_Accuracy_EC_SUB_SUB_CLASS.pdf"


plot(performance(prediction(combine$RC, as.logical(combine$ECX) & as.logical(combine$ECX.X) & as.logical(combine$ECX.X.X)), measure = "sens", x.measure = "spec"), lwd=2.5, col="cornflowerblue",
main="Sensitivity/Specificity Curve for Tanimoto scores vs. EC Sub-SubClass Matches")

plot(performance(prediction(combine$BC, as.logical(combine$ECX) & as.logical(combine$ECX.X) & as.logical(combine$ECX.X.X)), measure = "sens", x.measure = "spec"), lwd=2.5, col="DARKRED", add=TRUE)

plot(performance(prediction(combine$ST, as.logical(combine$ECX) & as.logical(combine$ECX.X) & as.logical(combine$ECX.X.X)), measure = "sens", x.measure = "spec"), lwd=2.5, col="darkorange", add=TRUE)

plot(performance(prediction(combine_BC_RC_0_05$BC_RC, as.logical(combine_BC_RC_0_05$ECX) & as.logical(combine_BC_RC_0_05$ECX.X) & as.logical(combine_BC_RC_0_05$ECX.X.X)), measure = "sens", x.measure = "spec"), lwd=1.5, col="PURPLE", add=TRUE)

plot(performance(prediction(combine_BC_ST_0_05$BC_ST, as.logical(combine_BC_ST_0_05$ECX) & as.logical(combine_BC_ST_0_05$ECX.X) & as.logical(combine_BC_ST_0_05$ECX.X.X)), measure = "sens", x.measure = "spec"), lwd=1.5, col="BROWN", add=TRUE)

plot(performance(prediction(combine_RC_ST_0_05$RC_ST, as.logical(combine_RC_ST_0_05$ECX) & as.logical(combine_RC_ST_0_05$ECX.X) & as.logical(combine_RC_ST_0_05$ECX.X.X)), measure = "sens", x.measure = "spec"), lwd=1.5, col="yellow", add=TRUE)

legend("bottomright", c("Bond Changes (BC)","Reaction Center (RC)", "Structure Similarity (SS)", "BC+RC","BC+SS", "RC+SS"), lty=c(1,1,1,1,1,1), lwd=c(2.5, 2.5, 2.5,1.5, 1.5, 1.5), col=c("DARKRED","cornflowerblue", "darkorange","PURPLE","BROWN","yellow"))

quartz.save(filepng,type="pdf",device=dev.cur(),dpi=100)
dev.off()
```

REFERENCES:

1. Todd, A. E. A., Orengo, C. A. C. & Thornton, J. M. J. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology* **307,** 1113–1143 (2001).
2. O'Brien, P. J. P. & Herschlag, D. D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chem Biol* **6,** 0–0 (1999).
3. Arnold, F. H. & Georgiou, G. *Directed Enzyme Evolution*. (Springer, 2003).
4. Glasner, M. E., Gerlt, J. A. & Babbitt, P. C. Evolution of enzyme superfamilies. *Current Opinion in Chemical Biology* **10,** 492–497 (2006).
5. Khersonsky, O. & Tawfik, D. S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Biochemistry* – (2010). doi:10.1146/annurev-biochem-030409-143718
6. Nguyen, T. T. *et al.* The Mechanism of the Reaction Catalyzed by Uronate Isomerase Illustrates How an Isomerase May Have Evolved from a Hydrolase within the Amidohydrolase Superfamily. *Biochemistry* **48,** 8879–8890 (2009).
7. Furnham, N., Garavelli, J. S., Apweiler, R. & Thornton, J. M. Missing in action: enzyme functional annotations in biological databases. *Nat Chem Biol* **5,** 521–525 (2009).
8. Hofmann, B. *et al.* Structural investigation of the cofactor-free chloroperoxidases. *Journal of Molecular Biology* **279,** 12–12 (1998).
9. Furnham, N. N. *et al.* Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. *PLoS Comput Biol* **8,** e1002403–e1002403 (2012).
10. Cuff, A. L. *et al.* Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Research* **39,** D420–D426 (2010).
11. Lees, J. J. *et al.* Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Research* **40,** D465–D471 (2012).
12. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research* **41,** D43–D47 (2013).
13. Gasteiger, J. *Handbook of chemoinformatics*. (Vch Verlagsgesellschaft Mbh, 2003).
14. Faulon, J.-L. & Bender, A. *Handbook of Chemoinformatics Algorithms*. (Chapman and Hall/CRC, 2010).
15. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J Cheminform* **5,** 7 (2013).

16. Jochum, C., Gasteiger, J. & Ugi, I. The Principle of Minimum Chemical Distance(PMCD). *Angew. Chem. Int. Ed. Engl.* **19,** 495–505 (1980).