Supplementary Data

Common preparation procedures

Table 1SD. Properties of molecular hash codes.

Molecular HASH codes	Tautomer invariant	Stereo sensitive
HASHY	no	no
HASHSY	no	yes
TAUTO_HASH	yes	no
STEREO_TAUTO_HASH	yes	yes

Table 2SD. Complete overview of the computational preparation steps for each chemical vendor collection.

Chemical provider	Records in the original file	Failed 3D conversion	After Corina	Failed hash code calculation	Internal duplicates[a]	After internal duplicates removal
Asinex	261606	21	261585	0	208	261382
Chembridge	603253	331	602922	0	352	602575
ChemDiv	789834	24	789810	0	27618	762192
Life Chemicals	316539	2	316537	0	276	316261
NCI Open	260071	2523	257548	171	17450	239927
Princeton	1082074	34	1082040	32	107324	974684
Enamine	1350112	12	1350100	28	18	1350054
Specs	201570	14	201556	0	1	201555

a) Internal duplicates include stereoisomeric forms that are removed from original chemical provider databases. This information is restored in subsequent preparation steps where stereisomers are recalculated more thoughtfully

Table 3SD. Overview of the tautomer and stereoisomer preparation steps.

Cumulative records after file cleaning and duplicate removals	Added tautomers	Added stereoisomers	Cumulative records after tautomers and stereoisomers addition
3727751	256838	2996967	6981566

CoCoCo-SC

Table 4SD. Example of a CoCoCo-SC sdf entry

Exemple of sdf record

```
951


 51 54  0  0  1  0            999 V2000
  -0.6884    1.7210   -0.3023 N   0  0  1  0  0  0  0  0  0  0  0  0
  -0.9539   -0.6067   -0.7657 N   0  0  1  0  0  0  0  0  0  0  0  0
  -0.2422    0.3652    0.0859 C   0  0  2  0  0  0  0  0  0  0  0  0
  -1.5114   -0.5940    5.6134 N   0  3  0  0  0  0  0  0  0  0  0  0
  -1.7043   -1.4951    0.0081 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.3645    2.6356   -0.2254 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.8260    0.1716   -1.6763 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.1635    1.5715   -1.7042 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.1848   -0.3472    4.1910 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.5747    0.1140    1.5342 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.8609   -1.3978    6.2570 O   0  5  0  0  0  0  0  0  0  0  0  0
  -2.4306    0.0067    6.1403 O   0  0  0  0  0  0  0  0  0  0  0  0
   0.4437    3.6839   -1.1334 C   0  0  0  0  0  0  0  0  0  0  0  0
  -2.9187   -1.0925    0.5492 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.9190    0.5698    3.4623 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.1416   -1.0291    3.5930 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.6140    0.8003    2.1339 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.1634   -0.7986    2.2645 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.4860    4.5863   -1.0545 C   0  0  0  0  0  0  0  0  0  0  0  0
  -3.6580   -1.9734    1.3139 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.3316    2.5022    0.7628 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.2404   -2.7849    0.2337 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.3714    3.4082    0.8359 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.9848   -3.6607    0.9993 C   0  0  0  0  0  0  0  0  0  0  0  0
  -3.1906   -3.2549    1.5412 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.4503    4.4473   -0.0731 C   0  0  0  0  0  0  0  0  0  0  0  0
  -4.9754   -1.5374    1.9016 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.5731    5.7236   -2.0395 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.8329    0.2725   -0.0677 H   0  0  0  0  0  0  0  0  0  0  0  0
  -1.8318   -0.2701   -2.6727 H   0  0  0  0  0  0  0  0  0  0  0  0
  -2.8385    0.2342   -1.2774 H   0  0  0  0  0  0  0  0  0  0  0  0
  -1.8940    2.3420   -1.9509 H   0  0  0  0  0  0  0  0  0  0  0  0
  -0.3264    1.5910   -2.4020 H   0  0  0  0  0  0  0  0  0  0  0  0
  -0.3089    3.7928   -1.9003 H   0  0  0  0  0  0  0  0  0  0  0  0
  -3.2841   -0.0917    0.3723 H   0  0  0  0  0  0  0  0  0  0  0  0
  -2.7315    1.1055    3.9306 H   0  0  0  0  0  0  0  0  0  0  0  0
   0.4351   -1.7420    4.1635 H   0  0  0  0  0  0  0  0  0  0  0  0
  -2.1907    1.5132    1.5633 H   0  0  0  0  0  0  0  0  0  0  0  0
   0.9754   -1.3348    1.7960 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.2704    1.6911    1.4732 H   0  0  0  0  0  0  0  0  0  0  0  0
  -0.2989   -3.1026   -0.1895 H   0  0  0  0  0  0  0  0  0  0  0  0
   3.1237    3.3051    1.6039 H   0  0  0  0  0  0  0  0  0  0  0  0
  -1.6248   -4.6636    1.1751 H   0  0  0  0  0  0  0  0  0  0  0  0
  -3.7708   -3.9417    2.1395 H   0  0  0  0  0  0  0  0  0  0  0  0
   3.2644    5.1545   -0.0137 H   0  0  0  0  0  0  0  0  0  0  0  0
  -5.7781   -1.7588    1.1983 H   0  0  0  0  0  0  0  0  0  0  0  0
  -5.1519   -2.0727    2.8346 H   0  0  0  0  0  0  0  0  0  0  0  0
  -4.9490   -0.4654    2.0972 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.0363    6.5869   -1.6460 H   0  0  0  0  0  0  0  0  0  0  0  0
   2.6186    5.9874   -2.1984 H   0  0  0  0  0  0  0  0  0  0  0  0
   1.1271    5.4192   -2.9863 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  3  1  0  0  0  0
```

```
  1  6  1  0  0  0  0
  1  8  1  0  0  0  0
  2  3  1  0  0  0  0
  2  5  1  0  0  0  0
  2  7  1  0  0  0  0
  3 10  1  0  0  0  0
  3 29  1  0  0  0  0
  4  9  1  0  0  0  0
  4 11  1  0  0  0  0
  4 12  2  0  0  0  0
  5 14  2  0  0  0  0
  5 22  1  0  0  0  0
  6 13  1  0  0  0  0
  6 21  2  0  0  0  0
  7  8  1  0  0  0  0
  7 30  1  0  0  0  0
  7 31  1  0  0  0  0
  8 32  1  0  0  0  0
  8 33  1  0  0  0  0
  9 16  1  0  0  0  0
  9 15  2  0  0  0  0
 10 17  2  0  0  0  0
 10 18  1  0  0  0  0
 13 19  2  0  0  0  0
 13 34  1  0  0  0  0
 14 20  1  0  0  0  0
 14 35  1  0  0  0  0
 15 17  1  0  0  0  0
 15 36  1  0  0  0  0
 16 18  2  0  0  0  0
 16 37  1  0  0  0  0
 17 38  1  0  0  0  0
 18 39  1  0  0  0  0
 19 28  1  0  0  0  0
 19 26  1  0  0  0  0
 20 27  1  0  0  0  0
 20 25  2  0  0  0  0
 21 23  1  0  0  0  0
 21 40  1  0  0  0  0
 22 24  2  0  0  0  0
 22 41  1  0  0  0  0
 23 26  2  0  0  0  0
 23 42  1  0  0  0  0
 24 25  1  0  0  0  0
 24 43  1  0  0  0  0
 25 44  1  0  0  0  0
 26 45  1  0  0  0  0
 27 46  1  0  0  0  0
 27 47  1  0  0  0  0
 27 48  1  0  0  0  0
 28 49  1  0  0  0  0
 28 50  1  0  0  0  0
 28 51  1  0  0  0  0
M  CHG  2   4   1  11  -1
M  END
> <CoCoCo_ID>
951

> <VENDOR>
ANG,CHD

> <M_VENDOR_TAUTOMER>
1
```

```
> <M_TAUTO_HASH>
CF52BCE9F38F105B

> <M_HASHY>
CF52BCE9F38F105B

> <M_HASHSY>
CF52BCE9F38F105B

> <M_STEREO_TAUTO_HASH>
CF52BCE9F38F105B

$$$$
```

Table 5SD. Specifications of properties stored in the field codes of the CoCoCo-SC database

Field codes provided in the sdf file	Description
CoCoCo_ID	Integer that reports CoCoCo ID. The same integer serves also as a name for each sdf entry
VENDOR	String that reports in a comma separated format the abbreviations of the vendor that sells the entry molecules
M_VENDOR_TAUTOMER	If equals to 1 the reported tautomer is not calculated
M_HASHY	Molecular hash codes
M_HASHSY	Stereo sensitive molecular hash codes
M_TAUTO_HASH	Tautomer invariant molecular hash codes
M_STEREO_TAUTO_HASH	Tautomer invariant and stereo sensitive molecular hash codes

CoCoCo-Catalyst

Table 6SD. List of Catalyst parameters for conformational searches

Parameter	Description
Maximum conformers	250
Conformational sampling	FAST
Stereoisomer	Independent treatment of stereoisomers
Shapes	Inclusion of shape files
Properties files	Inclusion of SPST files

Table 7SD. Example of a chunk directory content for CoCoCo-Catalyst

Catalyst files	Description
CoCoCo.1.0.0bdb	Binary file that contains conformers
CoCoCo.1.0.0bdb.names	
CoCoCo.1.0.2bdb	Binary file that contains the 2D indexes, e.g. topologies and connection tables
CoCoCo.1.0.3bdb	Binary file that contains the 3D indexes, e.g. location constraints and hydrophobic points
CoCoCo.1.0.4bdb	Binary file that contains the shape indexes
CoCoCo.1.0.chm	ASCII file that contains feature dictionary
CoCoCo.1.bdb	ASCI file that contains the main configurations of the database, e.g. description of the locations of database components.
CoCoCo.1.spst	Text-delimited file properties and 1D data
fragments.data	Fragment database data file

Table 8SD. Detailed statistics of CoCoCo-Catalyst database chunks.

Chunk number	Number of molecules	Number of conformations
1	200000	11062882
2	200000	15622960
3	200000	16667837
4	200000	11157700
5	200000	15030250
6	200000	18809947
7	200000	23807970
8	200000	22879374
9	199999	16596423
10	199994	14741428
11	200000	19862243
12	200000	18985214
13	200000	18329070
14	200000	19088404
15	200000	23348086
16	200000	22089737
17	200000	17766190
18	200000	16843451
19	199994	15617406
20	199969	18285526

21	200000	20105689
22	200000	18144156
23	200000	17672499
24	200000	17699509
25	199996	17760788
26	200000	16669130
27	199421	14374772
28	196506	11716566
29	196356	13325512
30	199904	17977085
31	199775	18843606
32	199924	16962094
33	198868	18669685
34	199938	14398690
35	181555	13010398
Total	6972199	603922277

CoCoCo-Phase

Table 9SD. List of Ionizer parameters for calculating protonation states

Parameter	Description
pH of the medium	7.0
Structure treatment	Only the most favorable structure retained

Table 10SD. List of ConfGen parameters for conformational searches

Parameter	Description
Force field	OPLS 2005
Energy window	100 kJ/mol
Maximum number of conformers	250
Conformational sampling	Rapid
Conformer minimizations	100 steps
Stereoisomers	Independent treatment of stereoisomeric forms

Table 11SD. Example of a chunk directory content for CoCoCo-Phase

Catalyst files	Description
CoCoCo.1_dbInfo.log	Log file for Phase database management
CoCoCo.1_dbversion	Database version and format
CoCoCo.1_feature.ini	Feature definition file
CoCoCo.1_master_phase.inp	Ligand location file
CoCoCo.1_phasedb	Phase database file
CoCoCo.1_ligands	Directory that stores ligands allocated in sub-directory blocks of 5000 ligands each. The content of each block directory is a Phase binary file with extension h5

Table 12SD. Detailed statistics of CoCoCo-Phase database chunks.

Chunk number	Number of molecules	Number of conformations
1	199912	2738599
2	199963	3302655
3	199732	4027897
4	199812	2785118
5	199792	3360752
6	199718	4759352
7	199374	6669502
8	199509	6586037
9	199737	4030356
10	199825	3196729
11	199997	4443534
12	199969	4398318
13	199956	4644595
14	199982	4789208
15	200000	5278723

16	200000	4964345
17	200000	4480156
18	200000	4280742
19	199854	3512070
20	199958	4278353
21	199999	4765932
22	199998	4523082
23	199996	4306319
24	199998	4256723
25	199995	4339594
26	199975	4243287
27	197233	3472465
28	191030	2906590
29	193241	3398200
30	199640	3804865
31	199591	4320951
32	199652	3779748
33	198497	3779205
34	199619	3201917
35	181335	2854353
Total	6956889	144480272

CoCoCo-MC

Table 13SD. Detailed statistics of CoCoCo-MC database chunks.

Chunk number	Number of molecules	Number of conformations
1	199913	2738599
2	199964	3302655
3	199733	4027897
4	199812	2785118
5	199792	3360752
6	199718	4759352
7	199374	6669502
8	199509	6586037
9	199737	4030356
10	199825	3196729
11	199997	4443534
12	199969	4398318
13	199956	4644595
14	199982	4789208
15	200000	5278723
16	200000	4964345
17	200000	4480156
18	200000	4280742
19	199861	3512070
20	199960	4278353
21	200000	4765932
22	200000	4523082
23	199996	4306319
24	199998	4256723
25	199996	4339594
26	199975	4243287
27	197373	3472465
28	191097	2906590
29	193261	3398200
30	199640	3804865
31	199591	4320951
32	199652	3779748
33	198497	3779205
34	199621	3201917
35	181335	2854353
Total	6957134	144480272

CoCoCo databases electronic resources

URL: http://cococo.unimore.it

Contacts: Dr Alberto Del Rio
Dipartimento di Scienze Farmaceutiche
Università di Modena e Reggio Emilia
Via Campi 183
41100 Modena
Italy
Email mailto:alberto.delrio@unimore.it
Tel (+39) 059 2055161
Fax (+39) 059 2055131