

Supplementary Information

Development and Application of a Comprehensive Machine Learning Program for Predicting Molecular Biochemical and Pharmacological Properties

Hwanho Choi, Hongsuk Kang, Kee-Choo Chung, and Hwangseo Park*

*Sejong Battery Institute, Sejong University, 209 Neungdong-ro, Kwangjin-gu, Seoul 05006,
South Korea*

*Author to whom correspondence should may be addressed.

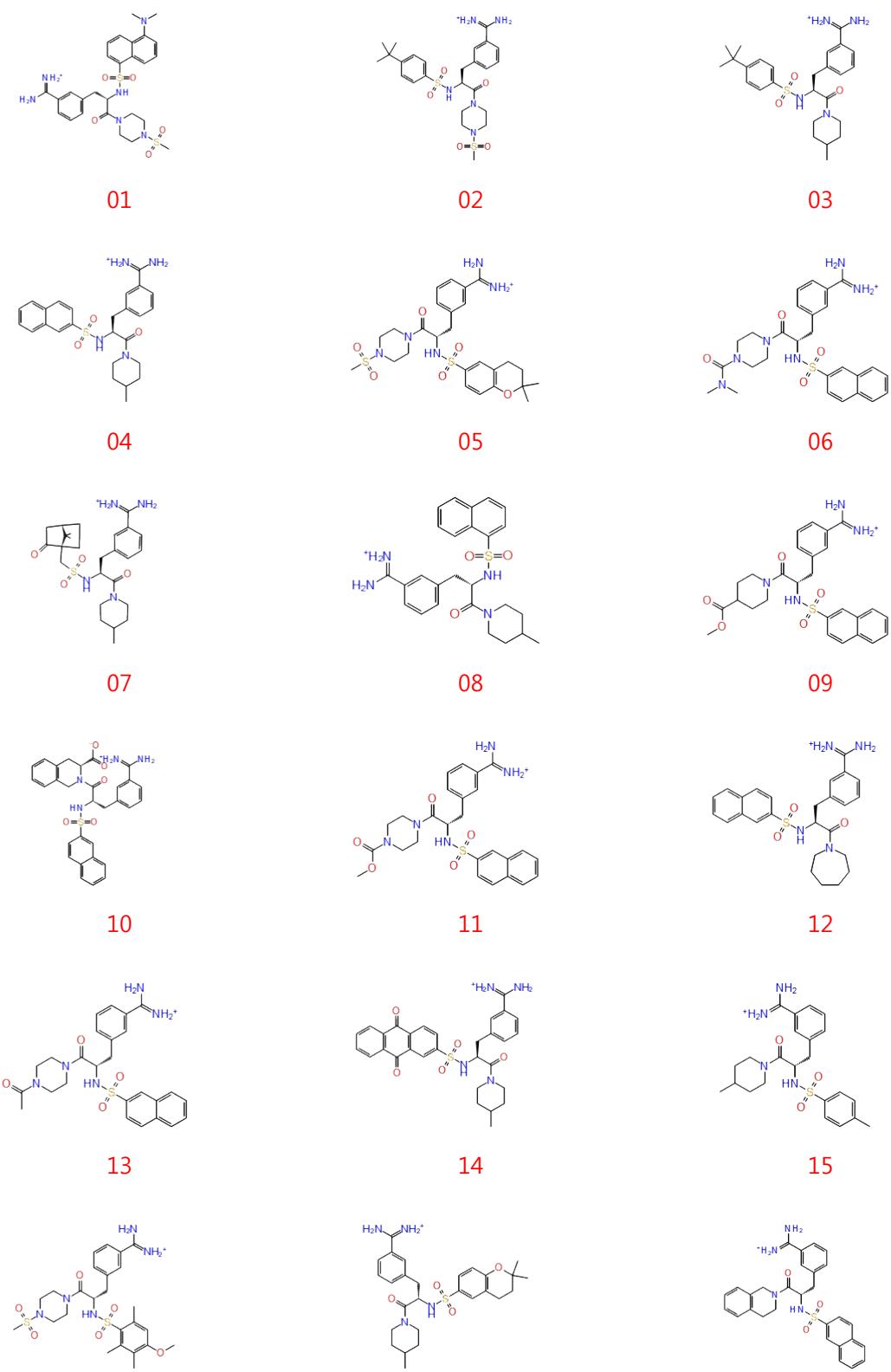
hspark@sejong.ac.kr

Chemical structures of the molecules in pK_i (thrombin), $\log D$, P_{Caco-2} ,
and CL_{int} datasets. S2-S16

Linear correlation diagrams between the experimental and calculated
 $\log D$ values of 53 SAMPL5 molecules in the 5-fold external cross-
validation. S17-S21

Linear correlation diagrams between the experimental and calculated
 $\log CL_{int}$ values of 37 drug molecules in the 5-fold external cross-
validation. S22-S26

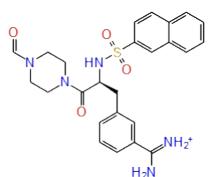
Table S1. Chemical structures of 88 thrombin inhibitors. Indicated in red are the molecules in the training set.



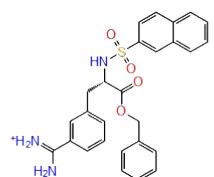
16

17

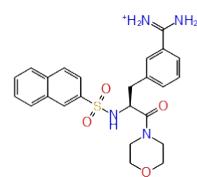
18



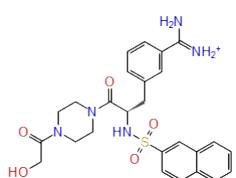
19



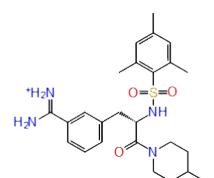
20



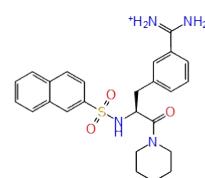
21



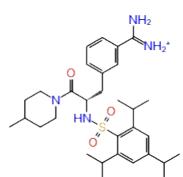
22



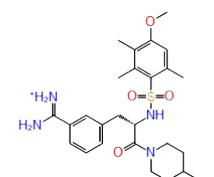
23



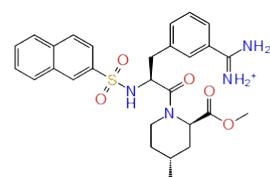
24



25



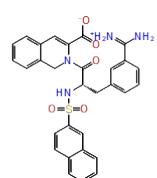
26



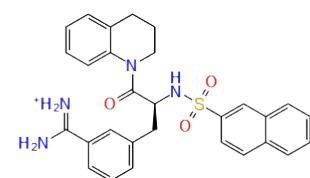
27



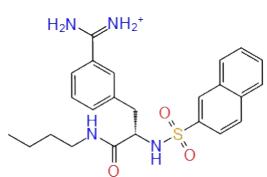
28



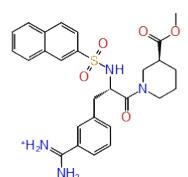
29



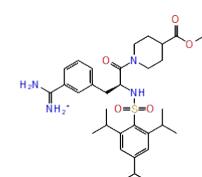
30



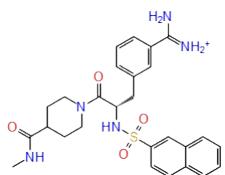
31



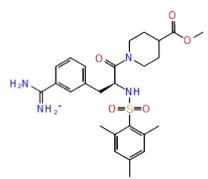
32



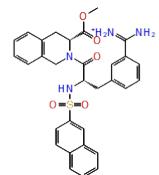
33



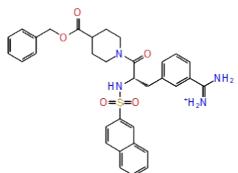
34



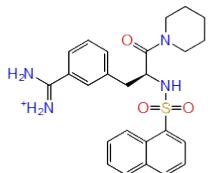
35



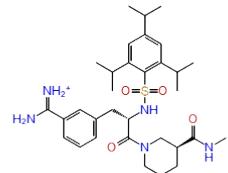
36



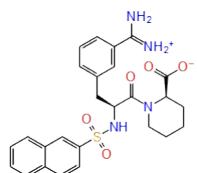
37



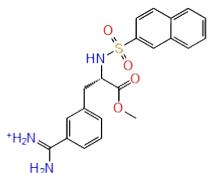
38



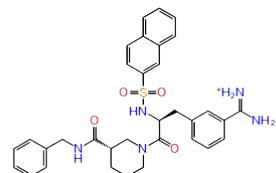
39



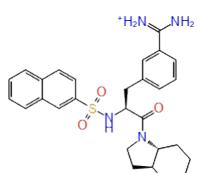
40



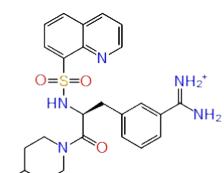
41



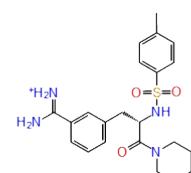
42



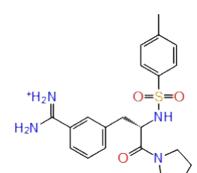
43



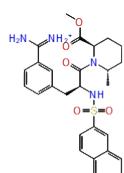
44



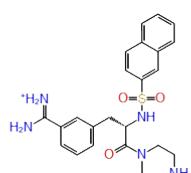
45



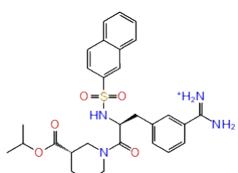
46



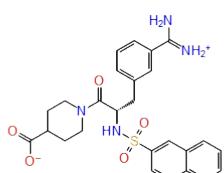
47



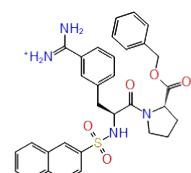
48



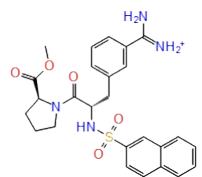
49



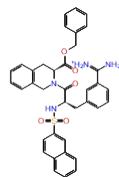
50



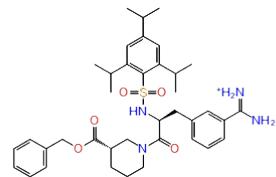
51



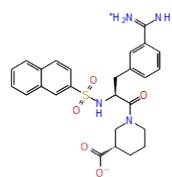
52



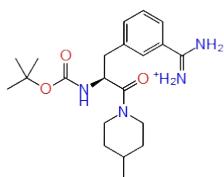
53



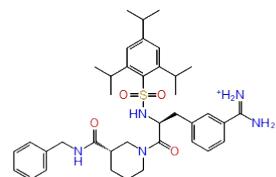
54



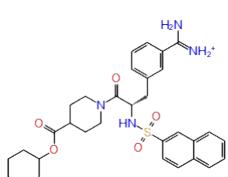
55



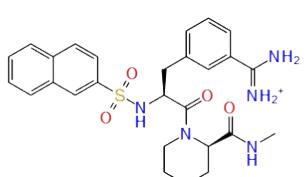
56



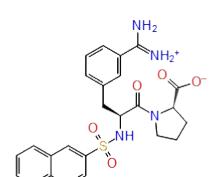
57



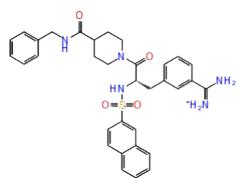
58



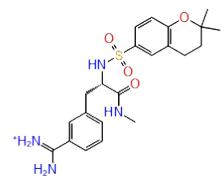
59



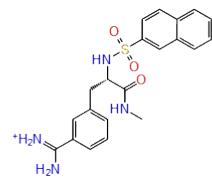
60



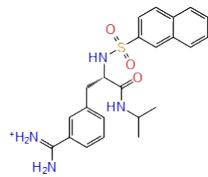
61



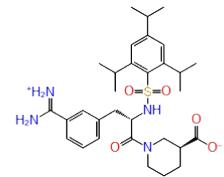
62



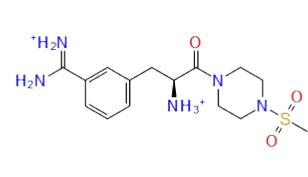
63



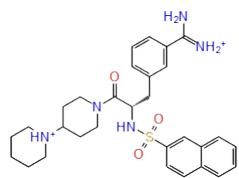
64



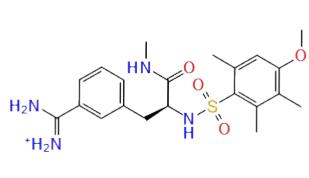
65



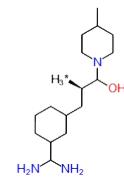
66



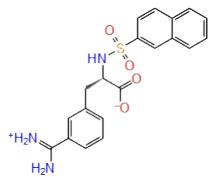
67



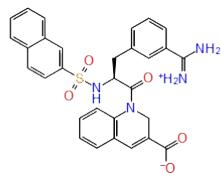
68



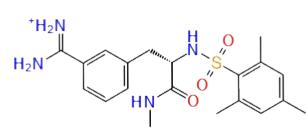
69



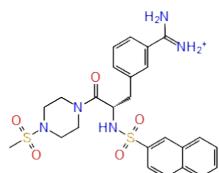
70



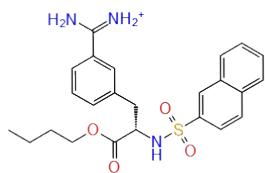
71



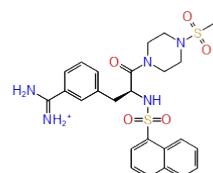
72



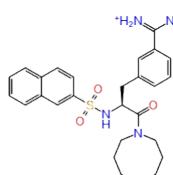
73



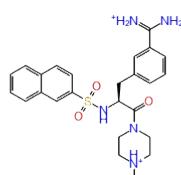
74



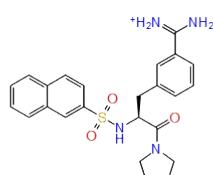
75



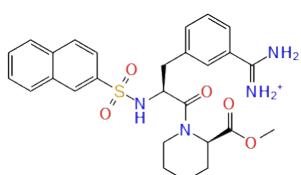
76



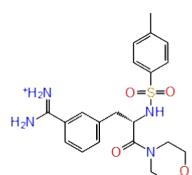
77



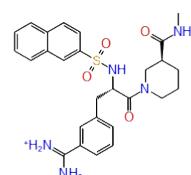
78



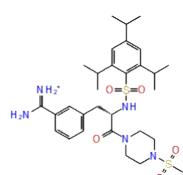
79



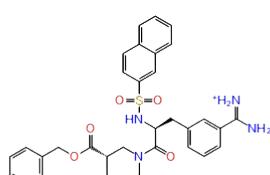
80



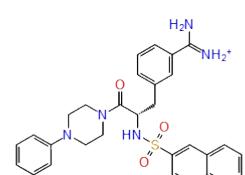
81



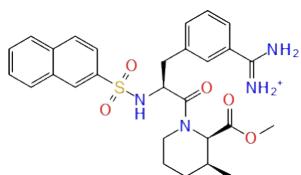
82



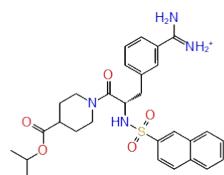
83



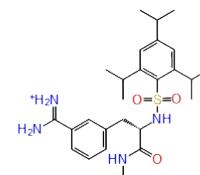
84



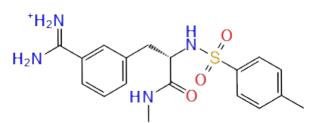
85



86

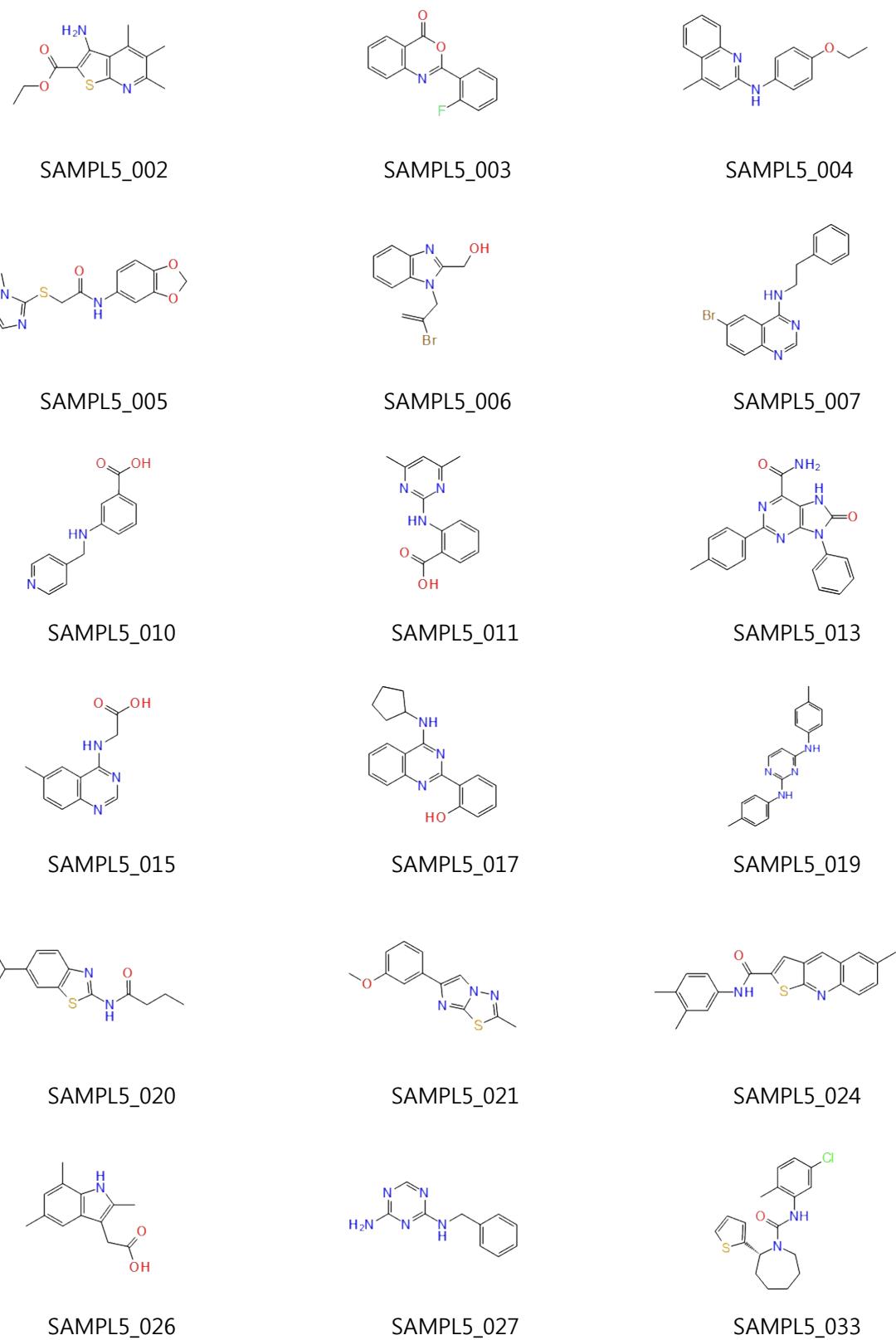


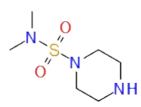
87



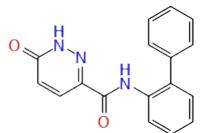
88

Table S2. Chemical structures of 53 SAMPL5 molecules

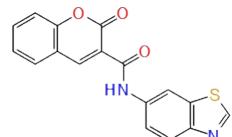




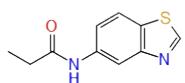
SAMPL5_037



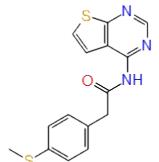
SAMPL5_042



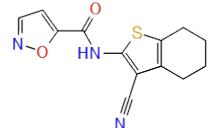
SAMPL5_044



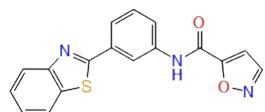
SAMPL5_045



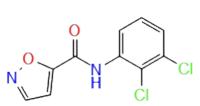
SAMPL5_046



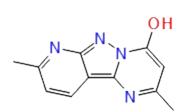
SAMPL5_047



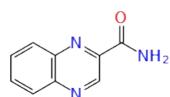
SAMPL5_048



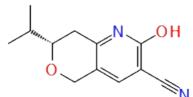
SAMPL5_049



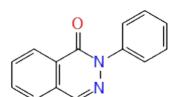
SAMPL5_050



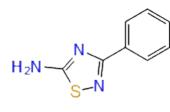
SAMPL5_055



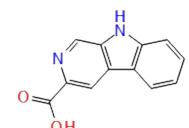
SAMPL5_056



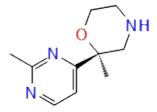
SAMPL5_058



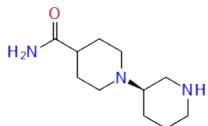
SAMPL5_059



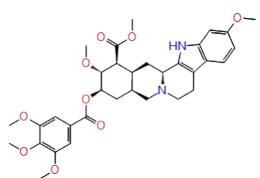
SAMPL5_060



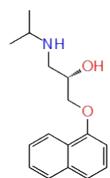
SAMPL5_061



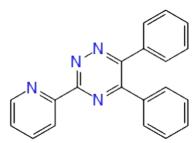
SAMPL5_063



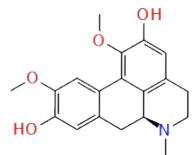
SAMPL5_065



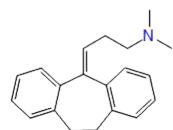
SAMPL5_067



SAMPL5_068



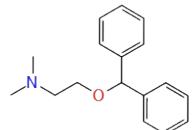
SAMPL5_069



SAMPL5_070



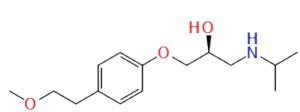
SAMPL5_071



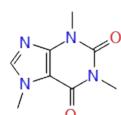
SAMPL5_072



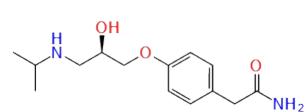
SAMPL5_074



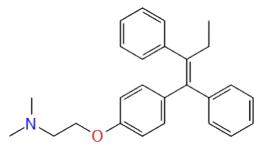
SAMPL5_075



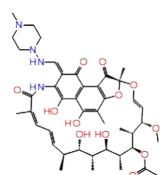
SAMPL5_080



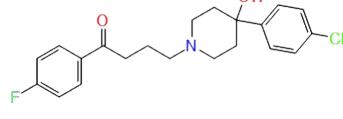
SAMPL5_081



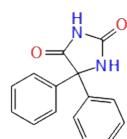
SAMPL5_082



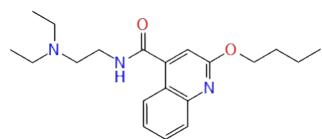
SAMPL5_083



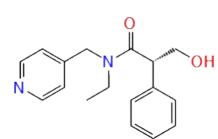
SAMPL5_084



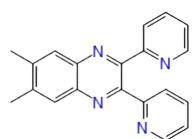
SAMPL5_085



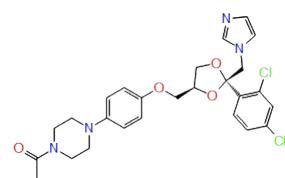
SAMPL5_086



SAMPL5_088

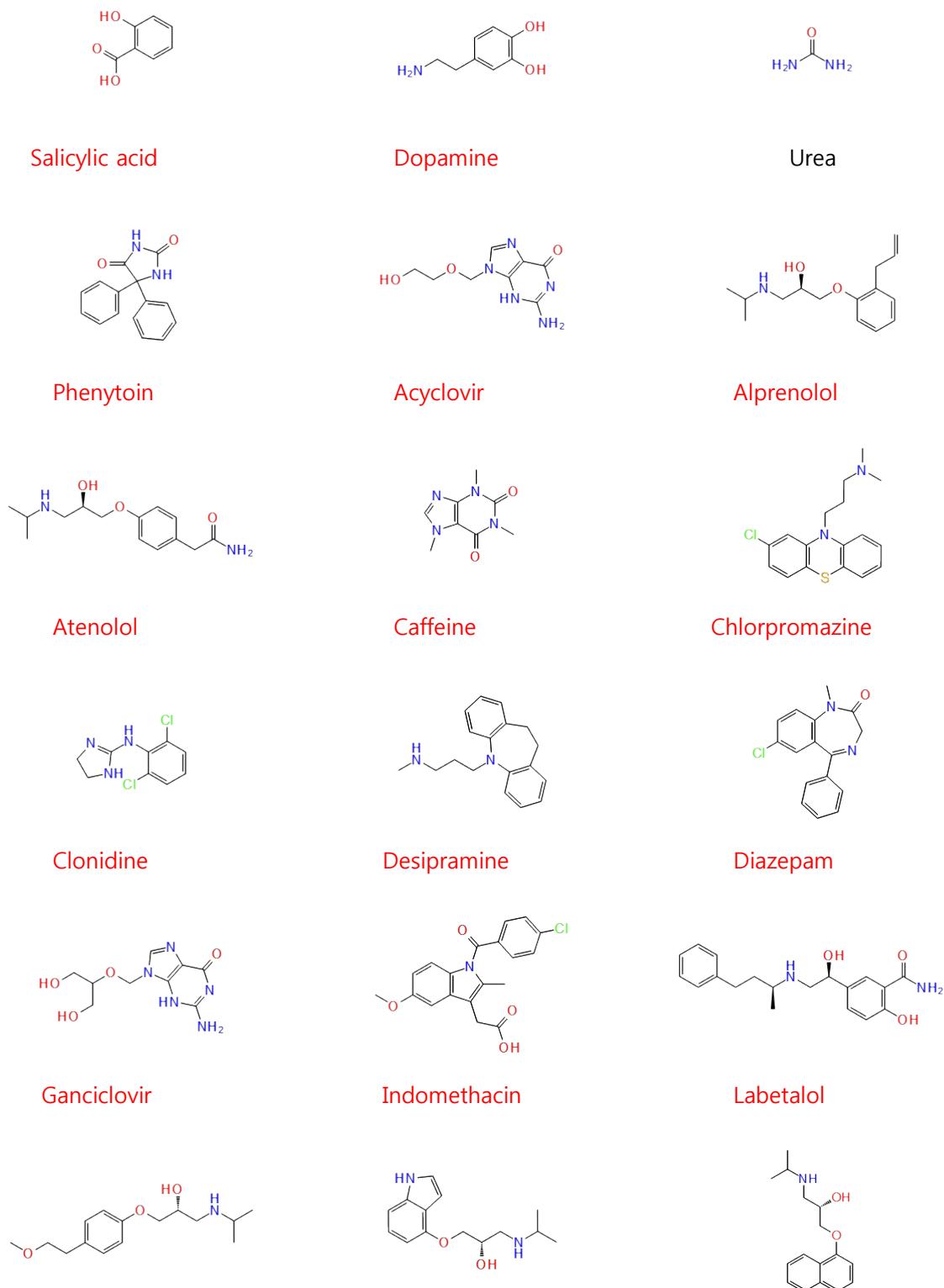


SAMPL5_090

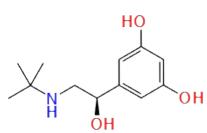


SAMPL5_092

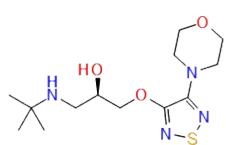
Table S3. Chemical structures of 38 drug molecules in Caco-2 dataset. Indicated in red are the molecules in the training set.



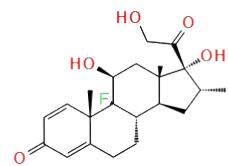
Metoprolol



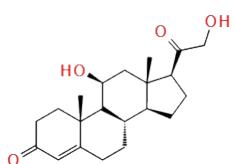
Pindolol



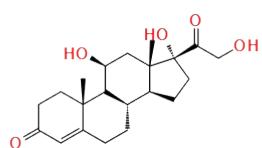
Propranolol



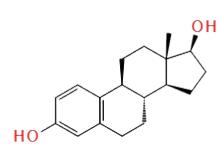
Terbutaline



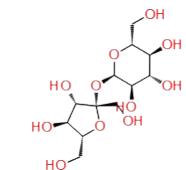
Timolol



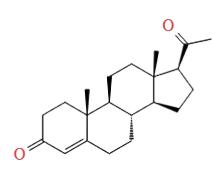
Dexamethasone



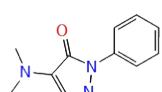
Corticosterone



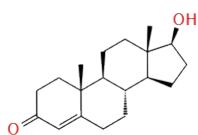
Hydrocortisone



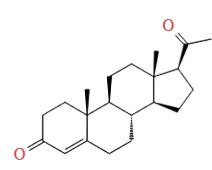
Estradiol



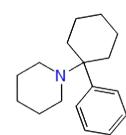
Sucrose



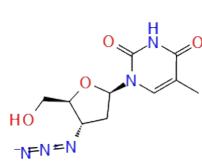
Progesterone



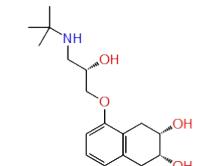
Aminophenazone



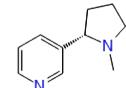
Testosterone



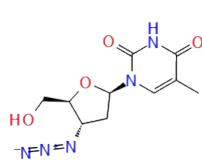
Mannitol



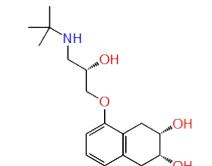
Phencyclidine



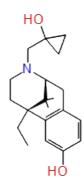
Zidovudine



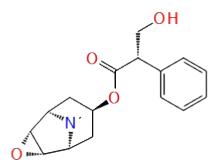
Nadolol



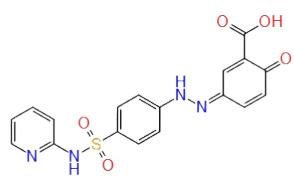
Nicotine



Bremazocine



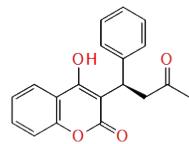
Scopolamine



Sulfasalazine

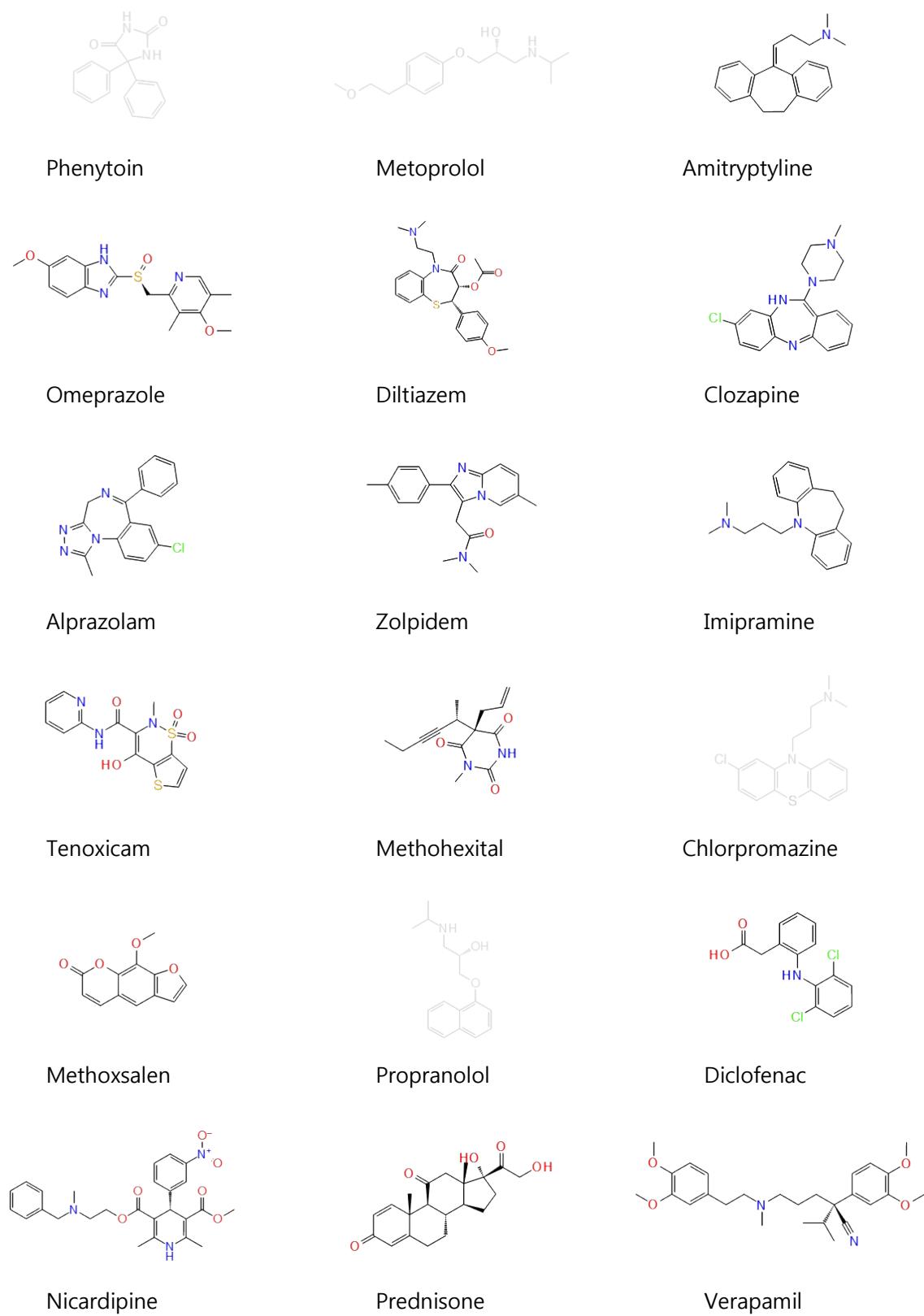


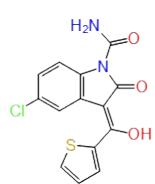
Meloxicam



Warfarin

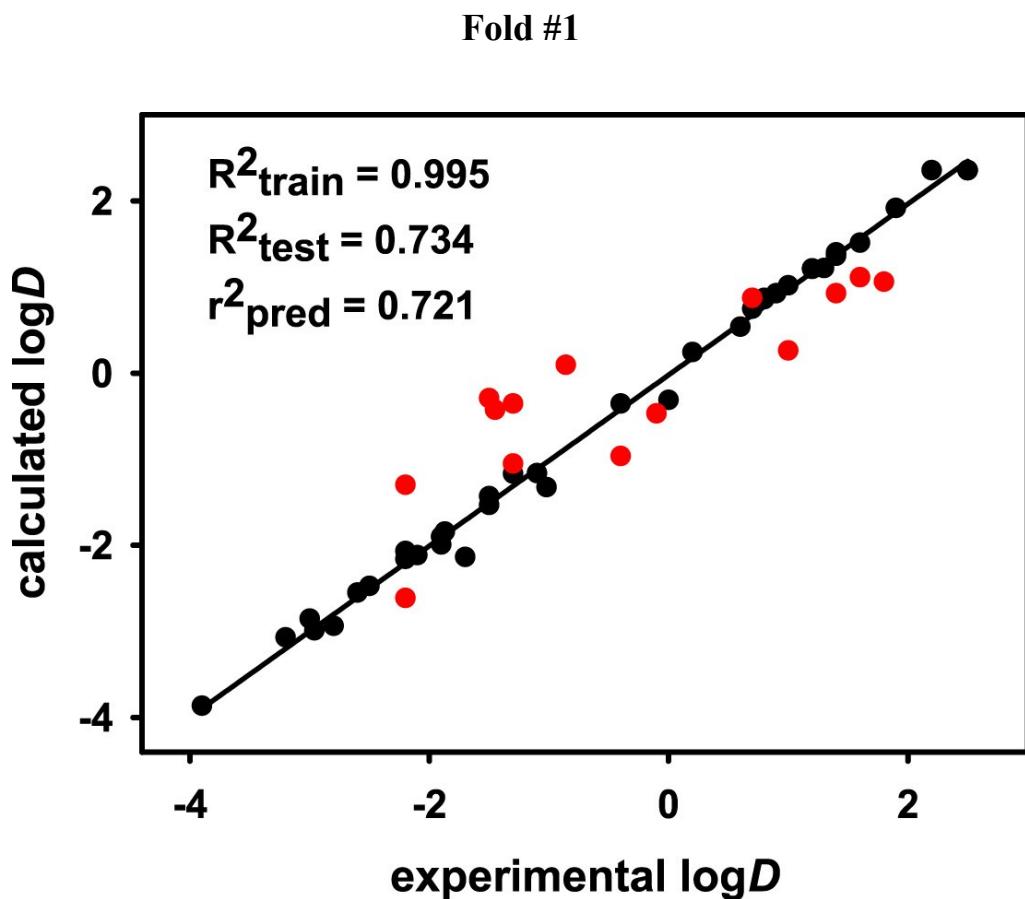
Table S4. Chemical structures of 37 drug molecules in CL_{int} dataset





Tenidap

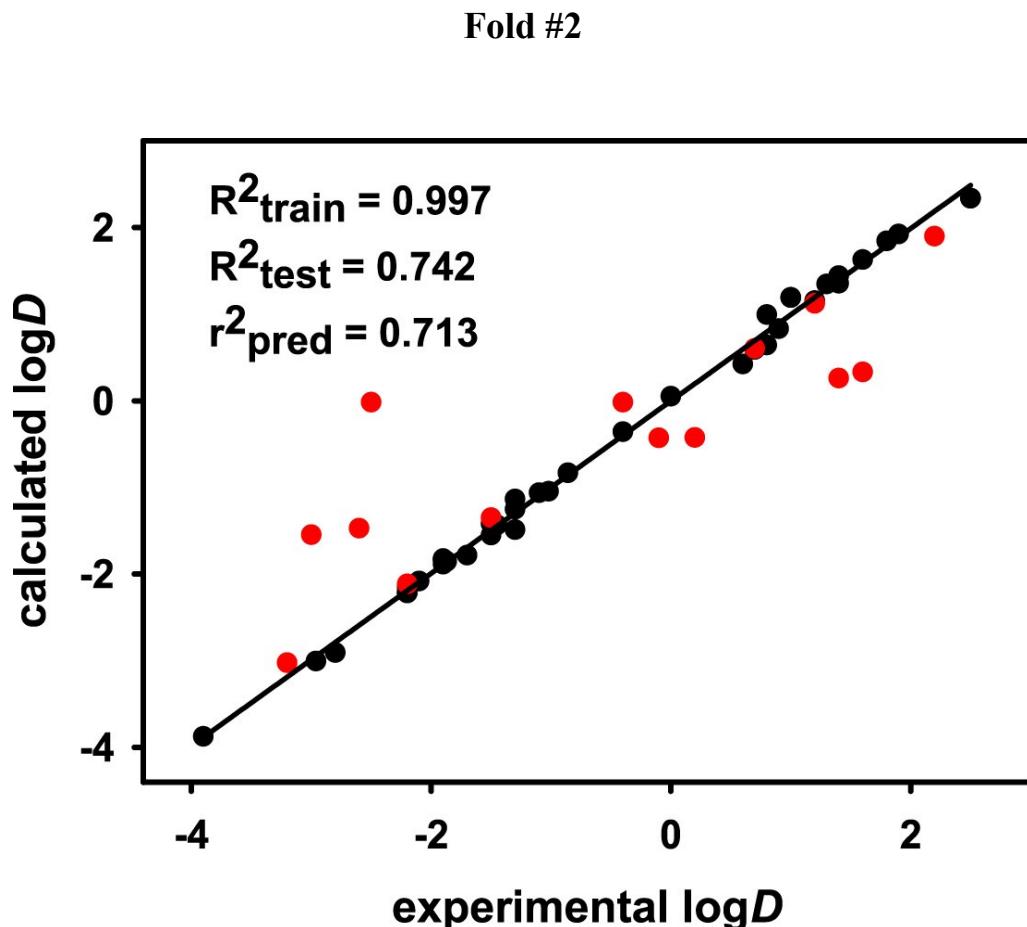
Figure S1. Linear correlation diagrams between the experimental and calculated results for $\log D$ values of the molecules in training (black) and test set (red).



Compounds in the training set: 002, 003, 004, 006, 007, 010, 011, 015, 017, 019, 021, 024, 026, 027, 037, 042, 045, 046, 048, 049, 050, 055, 056, 058, 059, 060, 063, 065, 070, 072, 074, 075, 081, 082, 083, 084, 088, 090, 092

Compounds in the test set: 005, 013, 020, 033, 044, 047, 061, 067, 068, 069, 071, 080, 085, 086

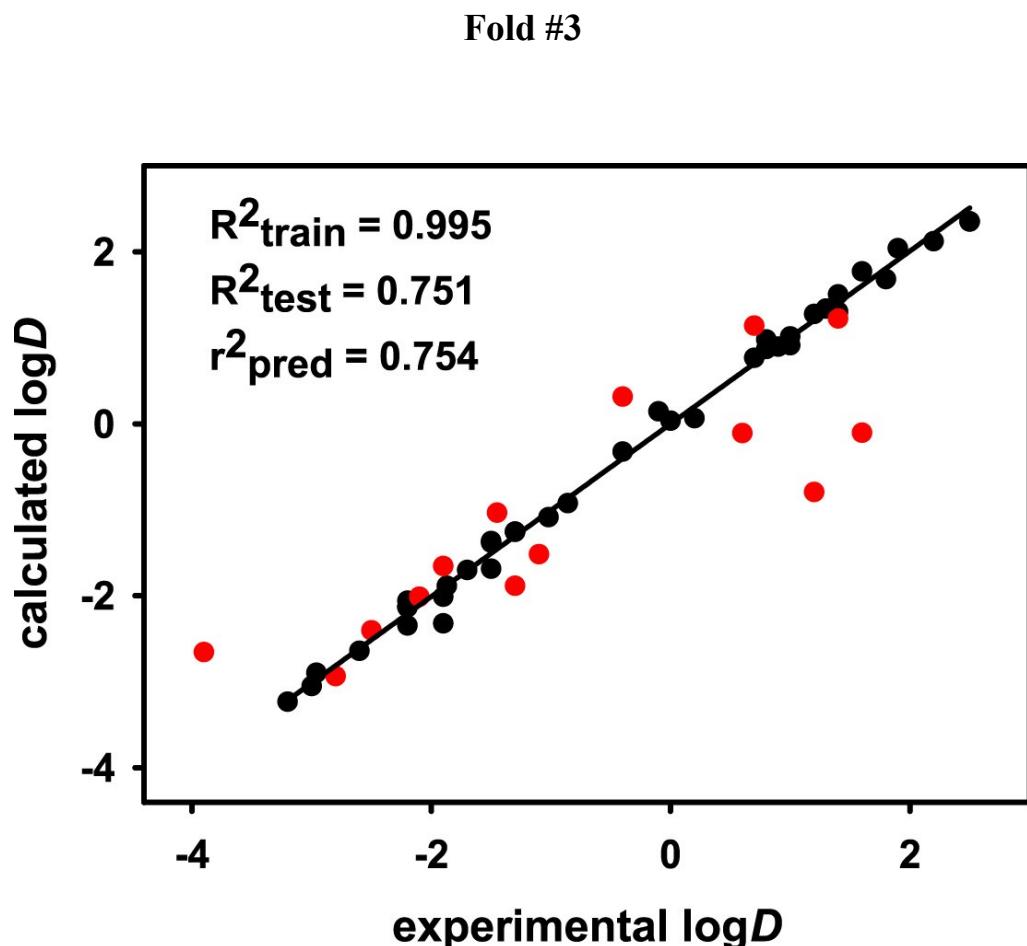
Figure S2. Linear correlation diagrams between the experimental and calculated results for LogD values of the molecules in training (black) and test set (red).



Compounds in the training set: 003, 005, 006, 007, 010, 011, 015, 017, 019, 020, 024, 027, 033, 037, 042, 044, 045, 048, 049, 055, 058, 059, 060, 061, 065, 067, 068, 069, 072, 074, 075, 080, 082, 083, 084, 085, 088, 090, 092

Compounds in the test set: 002, 004, 013, 021, 026, 046, 047, 050, 056, 063, 070, 071, 081, 086

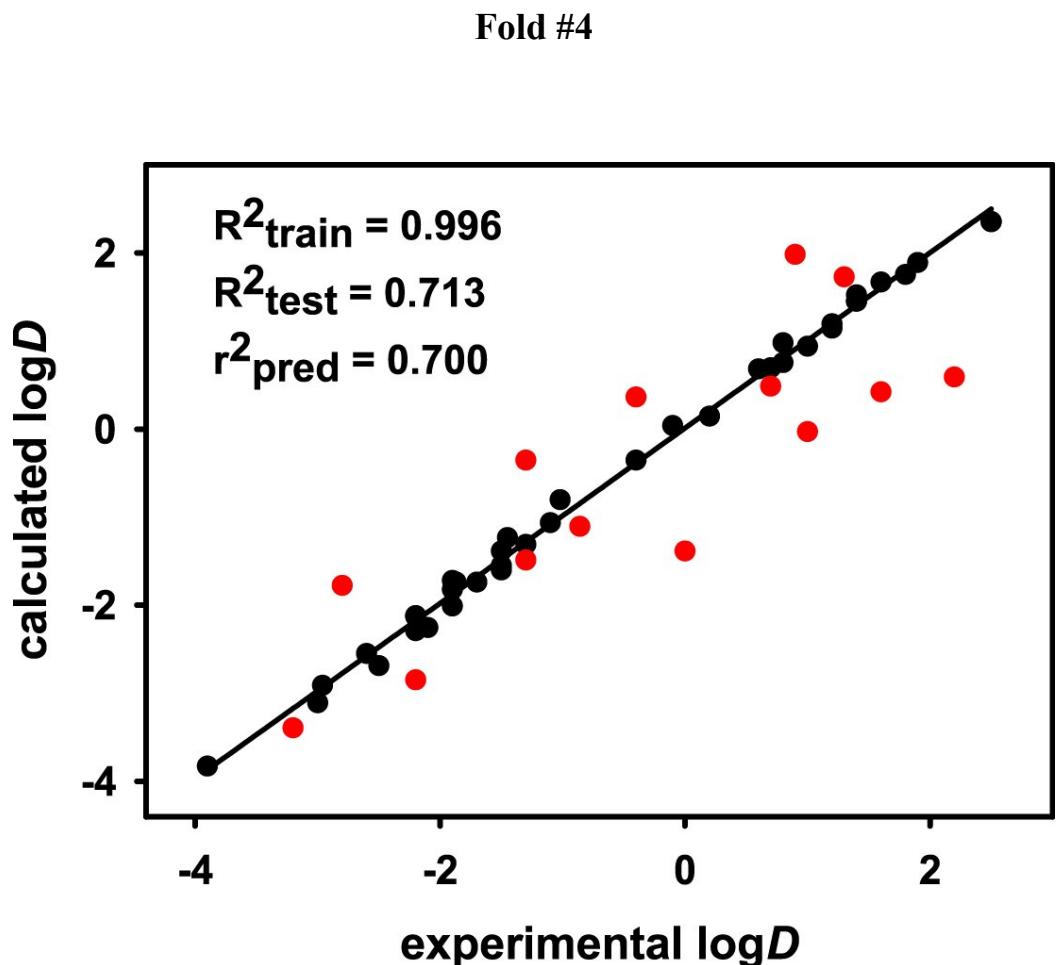
Figure S3. Linear correlation diagrams between the experimental and calculated results for LogD values of the molecules in training (black) and test set (red).



Compounds in the training set: 002, 003, 004, 005, 006, 007, 010, 011, 013, 015, 017, 020, 021, 024, 026, 027, 033, 037, 044, 046, 047, 048, 049, 050, 055, 058, 059, 063, 065, 069, 071, 080, 081, 082, 083, 084, 085, 088, 090

Compounds in the test set: 019, 042, 045, 056, 060, 061, 067, 068, 070, 072, 074, 075, 086, 092

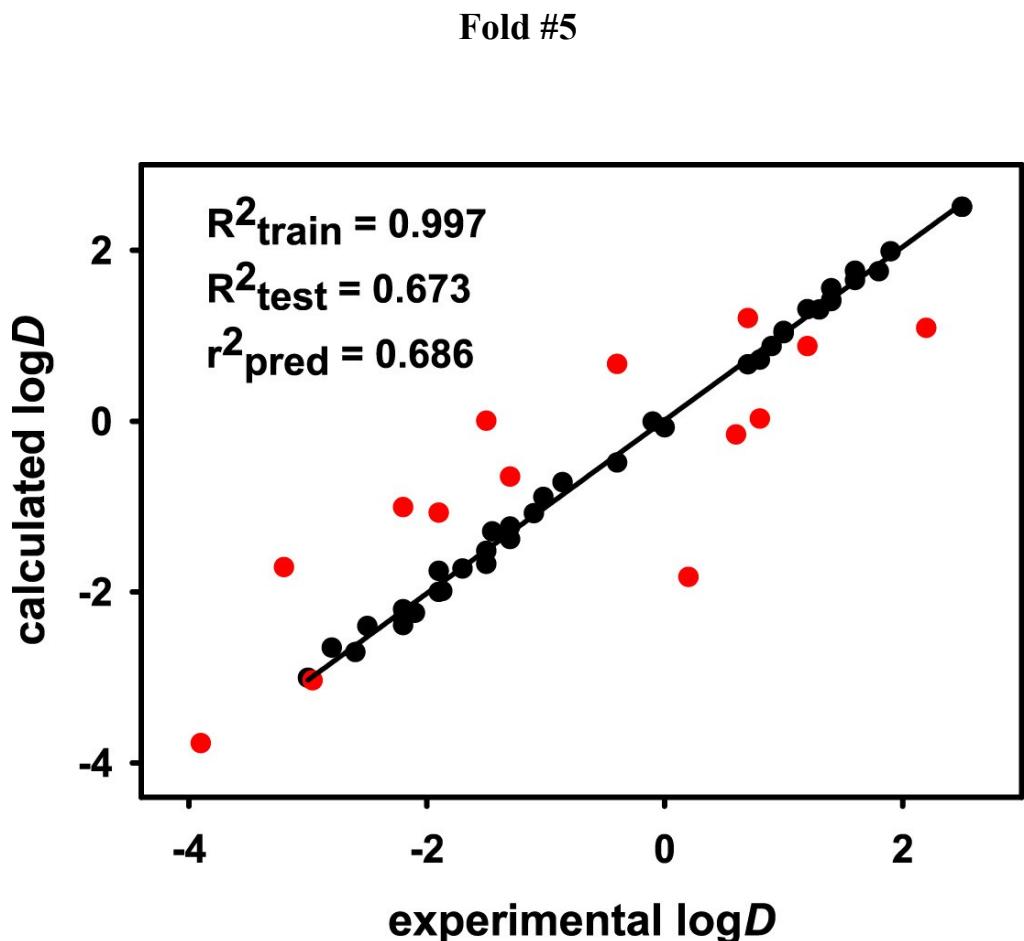
Figure S4. Linear correlation diagrams between the experimental and calculated results for LogD values of the molecules in training (black) and test set (red).



Compounds in the training set: 002, 003, 006, 007, 010, 011, 013, 015, 017, 019, 021, 026, 027, 033, 037, 042, 044, 045, 046, 047, 055, 056, 058, 059, 060, 061, 063, 068, 070, 071, 072, 074, 080, 082, 083, 085, 086, 088, 090

Compounds in the test set: 004, 005, 020, 024, 048, 049, 050, 065, 067, 069, 075, 081, 084, 092

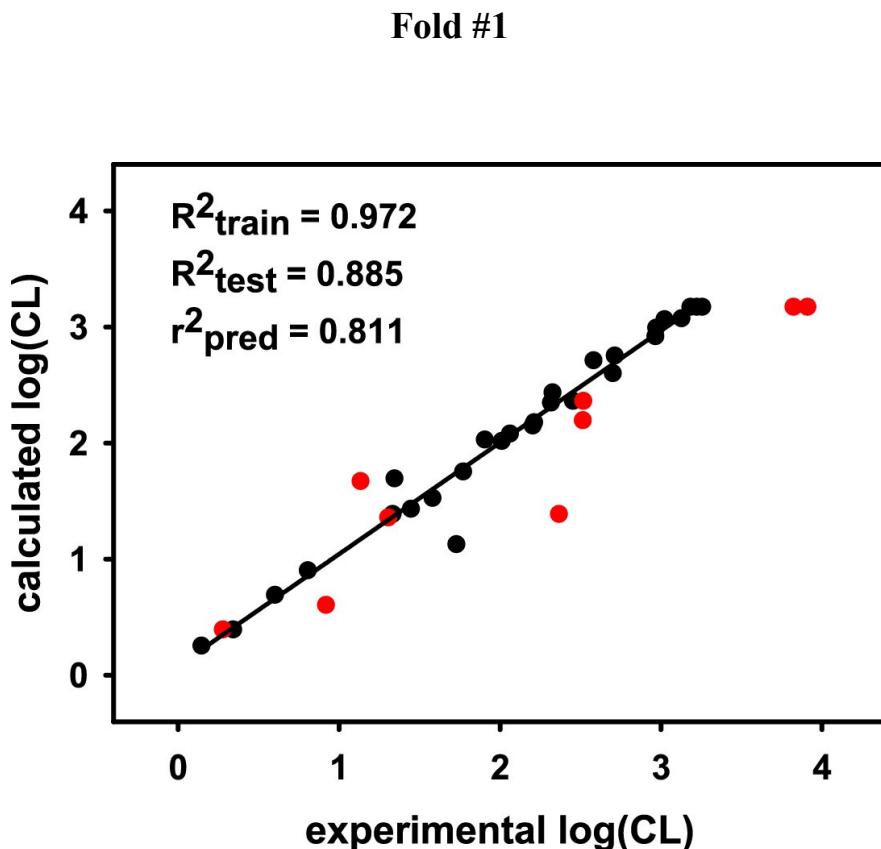
Figure S5. Linear correlation diagrams between the experimental and calculated results for LogD values of the molecules in training (black) and test set (red).



Compounds in the training set: 002, 003, 005, 006, 007, 010, 013, 015, 017, 019, 020, 024, 026, 027, 033, 037, 042, 044, 045, 048, 049, 056, 059, 061, 063, 065, 068, 069, 070, 071, 074, 075, 080, 082, 083, 084, 085, 090, 092

Compounds in the test set: 004, 011, 021, 046, 047, 050, 055, 058, 060, 067, 072, 081, 086, 088,

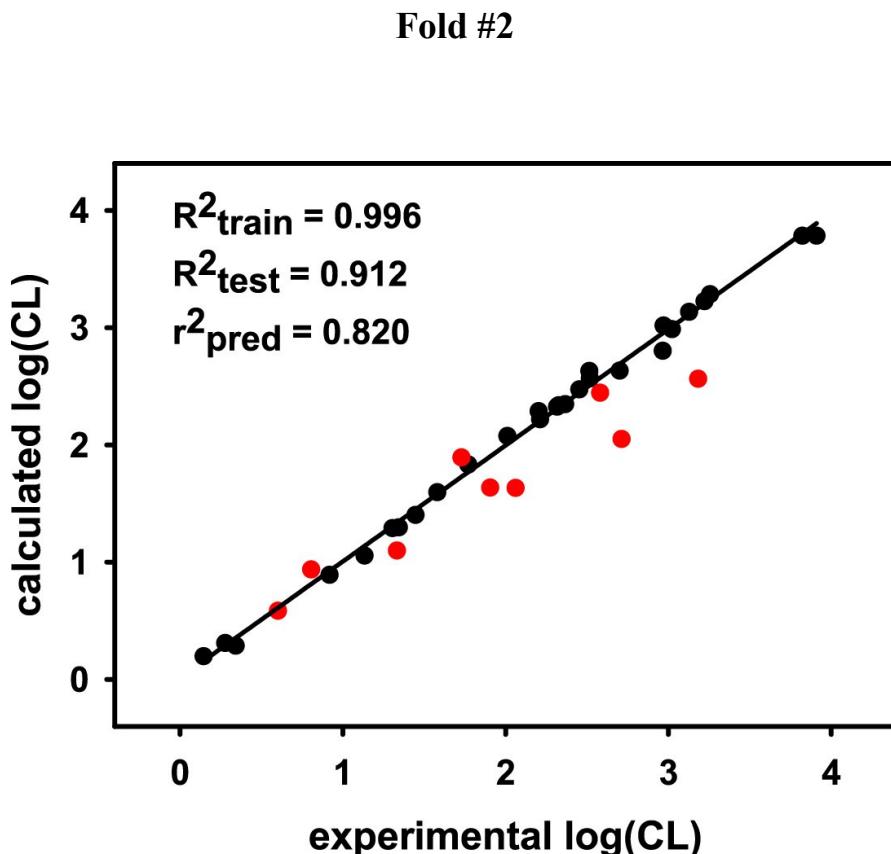
Figure S6. Linear correlation diagrams between the experimental and calculated results for logCL values of the molecules in training (black) and test set (red).



Compounds in the training set: phenytoin, amitryptyline, omeprazole, clozapine, zolpidem, tenoxicam, methohexital, chlorpromazine, methoxsalen, propranolol, diclofenac, nicardipine, prednisone, verapamil, midazolam, diazepam, triazolam, quinidine, ibuprofen, diphenhydramine, fluvastatin, tolbutamide, desipramine, amobarbital, phenacetin, fk1052, lorcaïnide, tenidap

Compounds in the test set: metoprolol, diltiazem, alprazolam, imipramine, propafenone, dexamethasone, hexobarbital, nilvadipine, fk480

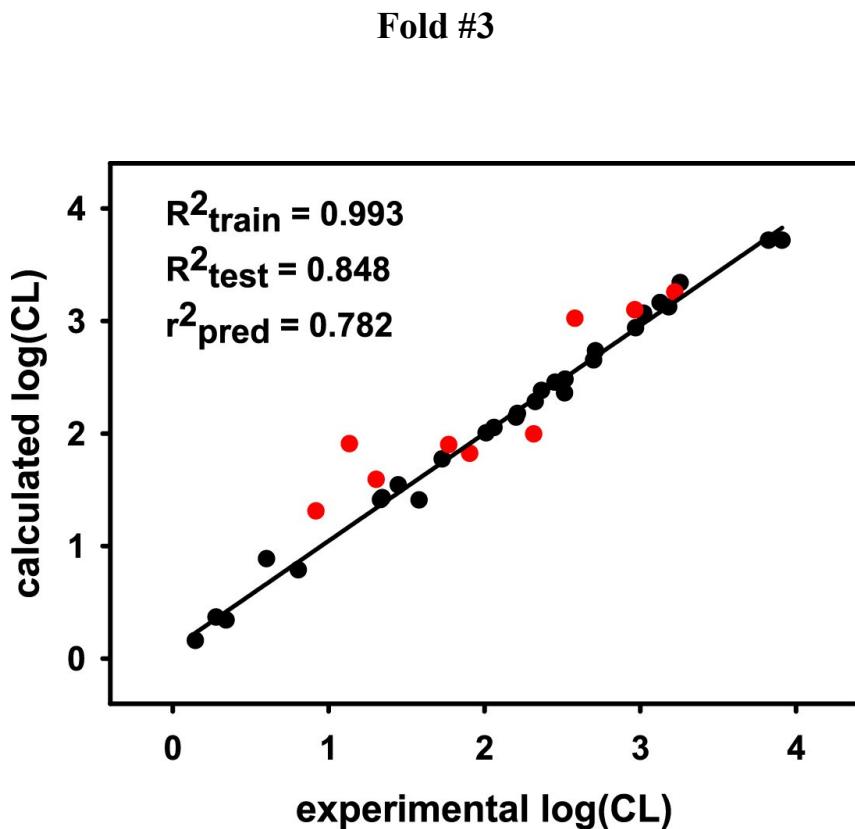
Figure S7. Linear correlation diagrams between the experimental and calculated results for logCL values of the molecules in training (black) and test set (red).



Compounds in the training set: metoprolol, omeprazole, diltiazem, clozapine, alprazolam, imipramine, tenoxicam, methohexital, methoxsalen, propranolol, diclofenac, nicardipine, verapamil, midazolam, diazepam, triazolam, quinidine, ibuprofen, fluvastatin, desipramine, propafenone, dexamethasone, amobarbital, hexobarbital, phenacetin, nilvadipine, fk480, lorcainide

Compounds in the test set: phenytoin, amitriptyline, zolpidem, chlorpromazine, prednisone, diphenhydramine, tolbutamide, fk1052, tenidap

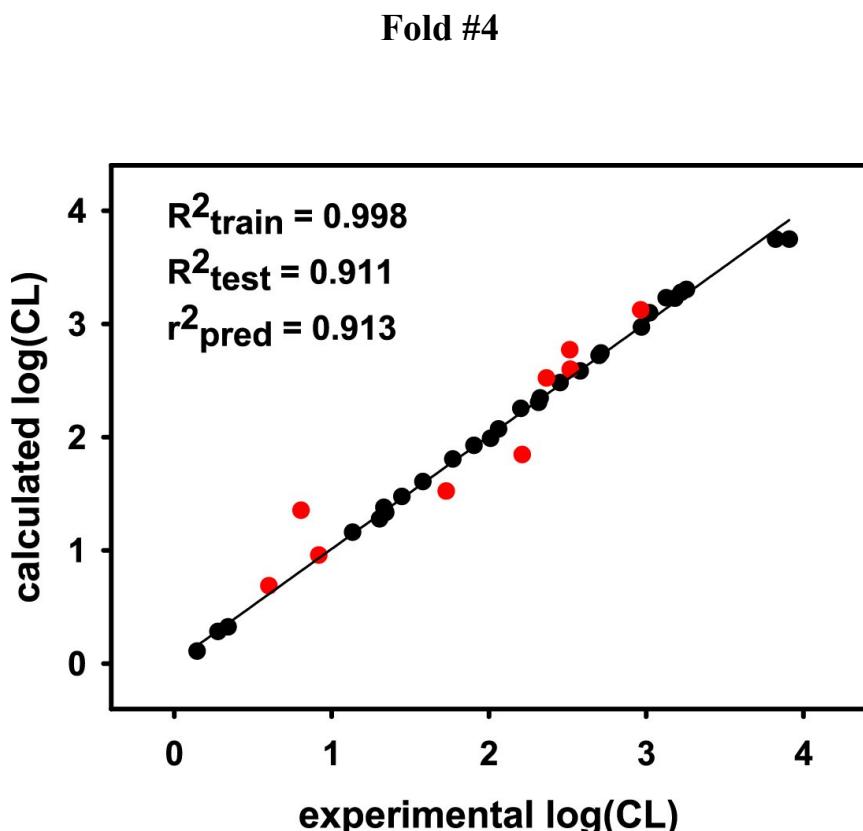
Figure S8. Linear correlation diagrams between the experimental and calculated results for logCL values of the molecules in training (black) and test set (red).



Compounds in the training set: phenytoin, amitryptyline, omeprazole, diltiazem, alprazolam, zolpidem, imipramine, tenoxicam, methoxsalen, propranolol, nicardipine, prednisone, verapamil, midazolam, diazepam, triazolam, quinidine, ibuprofen, diphenhydramine, fluvastatin, tolbutamide, desipramine, propafenone, amobarbital, phenacetin, nilvadipine, fk1052, fk480

Compounds in the test set: metoprolol, clozapine, methohexital, chlorpromazine, diclofenac, dexamethasone, hexobarbital, lorcaïnide, tenidap

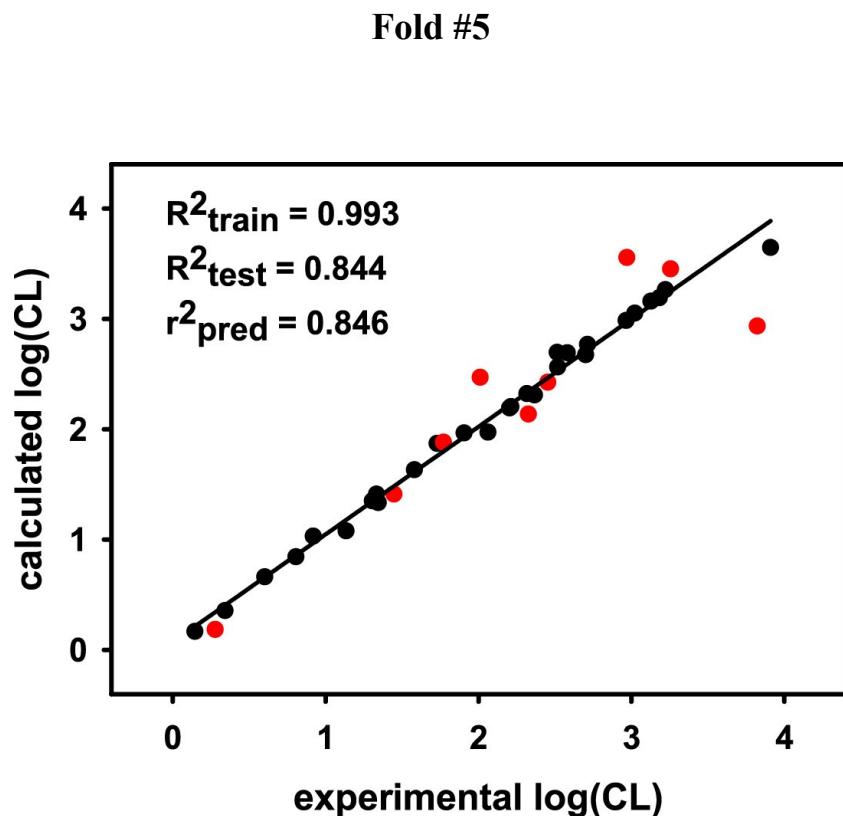
Figure S9. Linear correlation diagrams between the experimental and calculated results for logCL values of the molecules in training (black) and test set (red).



Compounds in the training set: metoprolol, amitryptyline, omeprazole, clozapine, alprazolam, zolpidem, tenoxicam, methohexital, chlorpromazine, methoxsalen, propranolol, diclofenac, nicardipine, prednisone, verapamil, diazepam, triazolam, quinidine, ibuprofen, fluvastatin, desipramine, propafenone, dexamethasone, amobarbital, phenacetin, nilvadipine, fk1052, tenidap

Compounds in the test set: phenytoin, diltiazem, imipramine, midazolam, diphenhydramine, tolbutamide, hexobarbital, fk480, lorcaïnide

Figure S10. Linear correlation diagrams between the experimental and calculated results for logCL values of the molecules in training (black) and test set (red).



Compounds in the training set: phenytoin, metoprolol, amitryptyline, omeprazole, diltiazem, zolpidem, imipramine, tenoxicam, methohexital, chlorpromazine, methoxsalen, diclofenac, prednisone, midazolam, triazolam, quinidine, diphenhydramine, fluvastatin, tolbutamide, desipramine, dexamethasone, amobarbital, hexobarbital, nilvadipine, fk1052, fk480, lorcaïnide, tenidap

Compounds in the test set: clozapine, alprazolam, propranolol, nicardipine, verapamil, diazepam, ibuprofen, propafenone, phenacetin