Index

1ACJ PDB file 359, 362–363, 365–367, 370
preparation of AChE from 366–367
1EVE PDB file 359, 366, 370
1-NN 277, 404
1OYT PDB file 315–322, 325, 330, 335, 341–343, 345
1ZGC PDB file 359, 366
2D autocorrelation vectors 130

a
A1A see alpha-1A adrenoceptor
A2A see human adenosine receptor
Accelrys 315
accuracy 164, 176
balanced (BA) 176
ACE see N-acetyl group
acetic acid 356
acetone 76
acetylcholine 356, 380, 390
N-acetyl group (ACE) 322
2-acetyl-5-methyl pyrrole 22
AchE see acetylcholinesterase
acridine scaffold 356
AD see applicability domain
AdaBoostM1 method 244
AdditiveRegression method 253–255
ADMET properties 291
alanine 71, 77
alanine amide 71
aliphatic hydrocarbons 285
alkanes 163, 164, 209
alpha-1A adrenoceptor (A1A) 119, 123, 124
Alzheimer’s disease 356
amber 62
Ambermax 62
amino acids 71, 286, 289, 314, 316, 319, 356
side chains 285
4-amino-5,6-dimethylthieno (2,3-d) pyrimidin-2(1H)-one 22
4,6-amino-triazyl 200
ANN see Artificial Neural Networks
anticoagulants 314
AOC see area over the curve
Apple 28
applicability domain (AD) 143–145, 153, 155, 160
aqueous solubility 249, 253, 263, 271
area over the curve (AOC) 197
area under the ROC curve (AUROC) 231, 237, 410
aromatic bonds 76, 419, 421
aromatic hydrocarbons 285
aromatic interactions 284, 286, 370
aromaticity 21, 76, 416, 420, 442
assigning 90
aromatic rings 22–23, 225, 289, 292, 359, 388
Artificial Neural Networks (ANN) 194
see also neural network
ASCII characters 75, 76
asparagine 315
aspartic acid 325, 390
atom-centred fragments 140
atoms 23, 76, 441–447
AUROC see area under the ROC curve

b
BA see balanced accuracy
backbone bond angles 320–321
backbone bond lengths 320–321
back-propagation algorithm 194
Back-Propagation Neural Network (BPNN) 210, 217
backtracking 433–436
bagging,
classification models 243–247
Random Forest method 264
regression models 249–255
balanced accuracy (BA) 176
Bayes formula 179
Benson’s notation 140
benzamidine 325, 330, 339
benzene 68, 71, 76, 325
BestFirst algorithm 165
binary fingerprints 79–80, 105–106
BioInfo database 243, 257, 263
biological activity 113, 119–124, 131, 332, 379, 402
bismuth oxychloride 22
bit collision 40
bit patterns 40
bit strings 40, 399
blood coagulation 314
blood-sucking insects 314
boiling points 163
Boolean operator 306
boosting,
classification models 243–247
regression models 249–255
bootstrapping 244, 250
bounding box 143, 153
BPNN see Back-Propagation Neural Network
branches 76
BRICS 100

C
caffeine 77, 175, 193
calcium oxide 22
CambridgeSoft 38
CANON algorithm 427–432
initial invariant 427
iteration step 428–431
canonicalization 421, 425–432
canonical SMILES 421–422, 425–432
canonic labels 77
carbamimidoyl group 323, 327–328, 330, 332–333, 335–337, 339, 348, 350
carboxybenzyl 198, 205
carboxylic acid 205, 286, 350
CAS number 4, 29
cation-π interactions 286
CBG see corticosteroid binding globulin
CCC see concordance correlation coefficient
CCG 38
CDC-like kinase (CDC) 86, 224–225, 231, 233, 236
CDK 395
Descriptor Calculator 119, 122, 129
CDK2 see cyclin-dependent kinase 2
central nervous system disease 356
cetareths 9
ceteths 9
ChemAxon 7, 41, 67, 75, 79, 395
ChEMBL 86, 175, 193, 292, 332, 381
ChEMBL20 396
ChemFinder 38
chemical cartridge 38
chemical databases,
data curation 3–36
distinguished from other databases 38
chemical functions 5, 23
Chemical Hashed Fingerprints (CHFP) 40–41
chemical ontology 23
chemical structures 4–5, 18, 21–24
aromatized 23
computed properties 41
dearomatized 23
standardization rules 46
storage in a database 38
substructures 40
chemical terms 41
chemogenomics 379
CHFP see Chemical Hashed Fingerprints
chiral atoms 76
chlorine 285
choline 356
class weighting 233
classification models 175
   algorithms 178–180
   bagging and boosting of 243–247
   step-by-step instructions 180–191
   theoretical background 176–178

see also compound classification

clustering,
   agglomerative 106–107
   average linkage 107
   complete-linkage 107
   divisive 106
   “good” 107
   k-means 106, 116
   methods 106
   procedure 97
   single-linkage 107

see also hierarchical clustering

coaugulation factor II 314

see also thrombin

coaugulation factor Xa (FXa) 119, 123

combinatorial enumeration 98–99

combinatorial libraries 67, 68, 98–101

compound classification,
   theoretical background 224–225
   using scikit-learn library 223–239

compound fragmentation 98–101

compound libraries 85–86
   analysis 90–95
   combinatorial libraries 98–101
   creation 87–90
   decorrelation 94–95
   descriptor calculation 90–94
   dimension reduction 94–95
   diverse subset selection 97–98
   normalization of descriptor values 91–92
   partitioning 95–97
   standardization 88, 91
   visualizing descriptor distributions 92–94

concordance correlation coefficient (CCC) 195

correlation coefficients 92, 142–143, 146, 155, 168, 171, 195, 219, 221

concordance (CCC) 195

leave-many-out (LMO) 143
leave-one-out (LOO) 143, 144, 152, 153, 155, 156
Matthews (MCC) 177–178

Pearson’s 144, 152, 156, 164, 177, 195
squared 150, 153, 155

correlation matrix 92–93
corticosteroid binding globulin (CBG) 129
Cosine coefficient 397
cost parameter 179
counterpropagation neural network (CPNN) 120, 121, 123–126

covalent bonds 88
COX2 see cyclooxygenase

CPNN see Counterpropagation Neural Networks
cross-docking 356, 366–370

cross-validation (CV),
   external 143, 164, 171–172
   internal 143, 164–171
   k-fold 225
   leave-many-out (LMO) 143, 145, 148
   leave-one-out (LOO) 143, 145, 148, 151, 157, 164

QSAR models 163–172

scikit-learn 233–234, 236, 238

crystallographic ligand 341–345
crystal water 356, 360
CSV format 46, 48, 51, 122, 180, 197

CXC CR4 antagonists 281
cyclic structures 76, 421
cyclin-dependent kinase 2 (CDK2) 119, 124, 289
cycholexanol 432
cychohexasiloxane 22
cychoxysinogenase (COX2) 119, 123
cysteine 325
d dabigatran 314
data,
   duplication 38
   erroneous 4
data (cont’d)
  frames 108
  fusion techniques 396
  mining 379
  quality 4
  separator 28
  structures 415–416
  valorization 3
  see also data curation
data curation,
  step-by-step instructions 7–34
  theoretical background 3–7
Daylight 395
  Daylight fingerprints 441
  Daylight theory manual 420
decision tree,
  compound classification 223, 225–226,
  231–234, 238
  visualization 232, 234
Derby 38
descriptors,
  1D 90
  2D 90
  3D 90, 130
calculation in compound libraries 90–94
decorrelation 94–94
dimension reduction 94–95
  hierarchical clustering 111–113
  ISIDA SMF 135, 139–142,
  179, 203,
  normalization of values 91–92
  RDF 130–132
  TLSER 209
  visualizing distributions 92–94
XLogP 131
Dice coefficient 403
dihedrals 320, 321
dimensionless numbers 113
diverse subset selection 97–98
docking 313
  benchmarks 355, 356
  retrospective virtual screening
  341–348, 350
  see also cross-docking; protein-ligand
  docking; re-docking
donepezil 356, 359, 366–370, 372
dopamine 380
  D3 receptor 381, 388, 390, 391
  D4 receptor 412
double bonds 76
drug profiling 281
drug repurposing 281
duplicate filtering 46
e
ECFP see extended connectivity
  fingerprints
E-DRAGON 129, 131–132
EL see Ensemble Learning
electrostatic interaction 327
enrichment factor 407
  Ensemble Learning (EL) 243–244, 250
EPA Fathead Minnow Acute Toxicity Database (EPAFHM) 129
EPAFHM see EPA Fathead Minnow Acute Toxicity Database
epsilon parameter (ε) 201
epsilon-insensitive loss function 194
epsilon-insensitive support vector machine
  (ε-SVM) 194
errors,
  classification 123, 180, 181, 185, 190,
  235, 244
  in databases 3, 12, 22, 29, 38
  during loading of data 48
  random 144
  regression models 194–196, 201, 202,
  204, 207, 252
  see also mean absolute error (MAE);
  regression error characteristic curve
  (REC); relative absolute error; relative
  mean square error; root-mean-square
  deviation (RMSD); root-mean-square
  error (RMSE)
S-ethyl 2-acetamidoethanethioate 22
eticlopside 381
Euclidean coefficient 78
Euclidean distance 97, 106, 111, 115, 120,
  215, 397
Euclidean space 106, 113
excluded volumes 284
extended connectivity fingerprints (ECFP)
  397, 399, 441–444
f
F1 score 237
False Negative Rate (FNR) 177
False Negatives (FNi) 177
False Positive Rate (FPR) 177, 178, 293
False Positives (FPi) 176, 407
feature lists 399
fibrin 314
fibrinogen 314
FILTER 381
fingerprints,
atom environment 441–447
binary 105–106, 399, 402
connectivity 106
count 397–400
Daylight 441
designs 106
encoding pharmacophore properties 106
extended connectivity (ECFP) 397, 399, 441–444
generation of 399–402
hierarchical clustering using 108–111
MACCS 24, 48, 105–108, 115, 397, 432
management of 40–41
molecular 24, 32, 397
Morgan 412 414
representations 399
virtual screening 397–402, 414–415
visualization 113, 115
Fischer’s criterion (F) 144, 145, 150, 152, 156
Fischer’s z-transformation 195
flexible alignment 334–335
FlexX 356, 360
fluorine 285
FNi see False Negatives
FNR see False Negative Rate
forward variable selection (FVS) 142–143, 155
FPR see False Positive Rate
fragment control 143, 153
Freedman paradox 164
FSMLR algorithm 142
FSN ligand 320, 323, 325–328, 333, 339, 341–343
furan 285
furan-2-yl-pyrazolo[3,4-d]pyrimidine 198, 205
FVS see forward variable selection
FXa see coagulation factor Xa
g
Gaussian functions 381
Gaussian probability density function 334
Gaussian Shape Similarity Score 305
GenerateMD program 79
generic structure 67
Gibbs free energy equation 285
Gini impurity 232, 233
Gini index 226
Givaudan 62
glutamic acid 325
glutamine 315
Good Scent Company 3
GOOGLE 4
G-protein coupled receptors 381
graph isomorphism problem 416, 427
graph theory 416–417
growing sets 258, 261
h
halogen 285
haloperidol 380–381, 387, 390
Hamilton R-factor percentage (HRF) 152, 156
hashed fingerprints 75–78
binary 79–80
generation 77–78
hash functions 443–444
heterocycles 285, 287
hierarchical clustering 105–117
algorithms 106–107
instructions 107–108
theoretical background 105–106
using descriptors 111–113
using fingerprints 108–111
visualization of the data sets 113–116
HierarchicalClusterPicker algorithm 97
Hill notation 419–420
hirudin 314, 315, 319, 320
hirudin peptide 320
histidine 315, 325, 365, 366, 372
HIV protease inhibitors 105, 107
Index

HRF see Hamilton R-factor percentage
Hückel’s rule 21
human adenosine receptor (A2A) 175, 180, 181, 185, 198
ligands 193, 194
hybridization 90
hydrocarbons 285
acceptors 284, 289, 326, 327, 332, 333
donors 284, 326, 327, 332, 333
hydrogen sulfide 22
hydrophobic interactions 284, 285
hydroxamates 287
“hypercubes” 96
hyperplanes 179, 227, 235–236
hypotension 380

I
Ibk method 163, 180, 216, 271
IDE see Integrated Development Environment
imidazole ring 315
InChIkey notations 4
InChl identifiers 75–80
canonicalization 77
generation of 77, 426
normalization 77
serialization 77
“induced fit effect” 355
insecticides 356
“instable” models 257
InstantJChem 37, 38, 40
adding new objects to database 62
administration interface 51, 53, 57, 59
customizing interface 60
data loading 48
Data Tree 43, 53, 59
database configuration 51
Database Tables 51
Database Views 51
Entities 43
field configuration 48
Fingerprints 46
ID Generation 46
installation 41–42
many-to-many relationships 56, 59–60
many-to-one relationships 53, 57–60
Odor Type table 51, 53, 55
Palette 60
Parsers option 51, 53, 55, 56, 57
Projects 43, 51, 53, 60
Schema 43, 51
Standardizer 46
Standard tables 43
Structure tables 43
Substance table 51, 57, 60, 62
Table creation options 51, 53, 55–57
Taste Reference tables 38, 57, 59, 60
trial licences 41
Views 43
Integrated Development Environment (IDE) 5
interpretable rules method 257, 261
step-by-step instructions 258–260
theoretical background 257–258, 261
ionic bonds 327
ionic interactions 286, 346
ISIDA/Model Analyzer 249, 257, 259
ISIDA/ModelAnalyzerR 193, 196, 200, 202–204, 263
ISIDA_QSPR 135–136
analysis of modeling results 144–148
applicability domain of the model 143–144
consensus models 153–161
Consensus Predictor tool 159–161
data split 139
EdiSDF tool 139
fitting equation 154–155
forward and backward stepwise variable selection 142–143
input data 137–139
internal model validation 143
model applicability domain 155
MolFrag module 149–150
n-fold cross validation 150–152, 155
property predictions 159–161
regression equations 142
RMSE estimation 148–151
setting the parameters 151
SMF descriptors 135, 139–142, 154–155
storage and retrieval of modeling results 144
variables selection 155
ISIDA SMF descriptors 135, 139–142, 179, 203, 204
isoceteths 9
IUPAC 76, 283
IUPHAR/BPS 175, 193
IupharDB 193

J
JATOON 119, 121, 123–125
JChem 75
cxcalc plugin 67–68, 72
standardizer 332
table 38, 40, 46
JChemExtension 7
JRip 258, 269

k
kernels 180, 194, 215, 227
keyword filtering 10
keywords 10
k-fold cross-validation 225
Kier-Hall topological indexes 209
k-Nearest Neighbors (k-NN) 163, 168–172, 176, 180, 187, 190, 191, 210, 215
stacking 271, 276–277
virtual screening 404, 407
KNIME 3, 34, 292, 395
Category to Number node 24, 26, 29
Cell Replacer node 22
Cell Splitter node 18
Column Combiner node 15, 16
Column Filter node 29
Column Merger node 12
Column Rename node 8–9, 22, 33, 34
Console 5
Create Bit Vector node 32
CSV Reader node 8, 12, 22
CSV Writer node 28, 34
data curation 3, 7–34
Domain Calculator node 12, 26
explorer 5
Favourite Nodes 5
GroupBy nodes 26, 28, 29
JChemExtension 7
Joiner node 12
MACCS Fingerprint node 24
Molecule Type Cast node 18
Naming node 24
node description 5
node repository 5
outline 5
overview 5–7
Row Filter node 9, 10, 22
Rule-based Row Filter node 10–11
Rule Engine node 11
shortcut buttons 5
Standardizer node 23
String Manipulation node 16, 18
String Replacer node 16, 18
Structure Checker node 18, 21
Table Writer node 21
workbench 5

k-NN see k-Nearest Neighbors
Kohonen self-organizing maps (Kohonen SOMs) 119–126
algorithms 120–121
applications 120
architecture of 120–121
building of neural network 121–124
instructions 121–126
theoretical background 119–120
Kubinyi fitness criterion 152, 156

l
Lagrange parameter 179
Lagrangian multipliers 227
lauryl ether sulfates 9
LBVS see ligand-based virtual screening
LeadIT 356–357, 359–360, 362
leave-many-out (LMO) cross-validation 143, 145, 148
leave-one-out (LOO) cross-validation 143, 145, 148, 151, 157, 164
correlation coefficient (Q^2) 143, 144, 152, 153, 155, 156
leeches 314
lepirudin 314
ligand-based pharmacophore models 298–301
ligand-based virtual screening (LBVS) 396
methods 397
search strategies 404
similarity metrics 402
ligand-protein complexes 294–298
ligand-receptor interactions 285–288
ligands,
acetylcholineesterase (AchE) 243, 257, 263
CDC-like kinase 4 223–224, 231
crystallographic 341–345
dopamine D4 receptor 412
input for docking 359–360
interactions with proteins 281
norepinephrine transporter 412
of human adenosine receptor (A2A) 193
profiling 379
serotonin 5HT1A receptor 412
shape constraints 287–288
lipophilic groups 285
linear notations 75
line ending character 28
Linux 28, 41, 296, 360, 387–390
Lipinski’s Rule of Five 332, 333
LMO see leave-many-out cross-validation
LOO see leave-one-out cross-validation
loop constructs 91
lysine 325

m
M5 algorithm 210
M5P see Regression Tree
MACCSC fingerprints 24, 48, 105–108, 115, 397, 400, 432
MACCSC keys 399, 441
machine learning methods 209, 271
“meta-learner” model 271–272
stacking 271–272
step-by-step instructions 210–221
theoretical background 209–210
MAE see mean absolute error
Maestro 38
magnesium oxide 22
many-to-many relationships 37
Markush, Eugene A. 68
Markush enumeration 67–68, 73
atom type variation 71
bond variation 71
chain length variation 71
drawing core structure 68
group position variation 71
library size estimation 68, 70
oligomers addition 71
step-by-step instructions 68–72
structure enumeration 70
MarkushSketch 68–70
Markush structure 67, 68, 70, 72
MarvinSketch 67, 68, 78
MarvinView 79
MASK file 139, 145, 148
matplotlib 85, 86, 395
imshow function 92
scatter function 93
scatterplots 93
Matthews correlation coefficient (MCC) 177–178
MaxMin algorithm 97
MaxMinPicker algorithm 97–98
MCC see Matthew’s Correlation Coefficient
MCL1 224, 233, 236
MDL SDfiles 78, 131
MDS see multidimensional scaling
mean absolute error (MAE) 144, 145, 187, 196
Merck Molecular Force Fields 325
metal binding interactions 287
metal complexation 286–287
metal ions 88, 284, 286–287
metalloproteases 286
metal oxides 22
methane 71
Microsoft 28
MLR see multiple linear regression
ModelAnalyzerR 193, 196, 200, 202–204, 206
MOE (Molecular Operating Environment) 38, 105–106, 315, 316, 320–337, 348, 395
database viewer (DBV) 332–333, 348
docking 341–345
London dG function 342–348
molar refractivity 90, 105
molecular descriptors 129–130
  algorithms 130–133
  classification 130
  generation of 129–134
molecular fingerprints 24, 32, 397
molecular representation 397
molecular structures,
  biological activity of 121–124
  representations of 75, 76
molecular weight 41, 48, 90, 94, 105, 112, 129, 332–333
  low 313, 320, 355, 379
Morgan algorithm 426, 427
Morgan fingerprints 412, 414
multidimensional scaling (MDS) 115
multinomial distribution 179
multiple linear regression (MLR) 135, 136, 142, 144, 149–150, 166, 210
  bagging procedure 249–252, 255
  stacking 271, 274, 275
MyChem 38
myocardial infarction 314
MySQL 38

n
Naïve Bayes (NB) 176, 178–179, 185, 187, 190, 191
  Bernoulli 230
  compound classification 223, 225, 230–231, 238–239
  Gaussian 230
  multinomial 180, 230
natural extracts 10
NB see Naïve Bayes
N-CA-C angles 321
neural network 194, 203–207
  counterpropagation (CPNN) 120, 121, 123–126
neurons 119, 123
  winning 120–121
n-fold cross validation,
  ISIDA_QSPR 150–152, 155
  QSAR models 163–164
  stacking 272
nitrogen,
  heterocycles containing 287
  pentavalent 23
  tetravalent 23
nitro group 23
NME see N-methyl amide
N-methyl amide (NME) 322, 325
n-NN 404
norepinephrine transporter ligands 412
n-propanol 76
numerical descriptors 105
numerical predictions 125–126
NumPy 85, 86, 91, 223, 395, 401
  calculating Euclidean distances 97
  standardization 91
NuSVC 235

O
OCHEM (Online Chemical Modeling Environment) 129, 132
(R)-(-)-2-octanol 22
octanol/water partition coefficient 90, 105, 112
octoxynols 9
odor category 4
odor description 4, 10–12, 15, 16, 28, 37
odor type 3, 5, 10–15, 24, 26, 28–29, 31, 37–38, 51, 53
OEChem TK 384
opleths 9
OMEGA2 381
one-to-many relationships 38
Openbabel 395
OpenEye 384, 395
OpenSMILES 420
Oracle 38
OrChem 38
organoleptic description 5, 15, 16, 18, 31, 62
organoleptic properties 3–5, 9, 32, 34
organoleptic semantic 5, 18
organoleptic substances 37–38
ortho-thiocresol 22
OSPR see quantitative structure-property relationship

P
PAM see Partitioning Around Medoids
paracetamol 78
pareths 9
Partial Least Squares (PLS)  210, 213
stacking  271, 274
partitioning  95–97
Partitioning Around Medoids (PAM)  116–117
patent claims  67
PCA see principal component analysis
PDB see Protein Data Bank
PDE5 see phosphodiesterase
Pearson’s correlation coefficient (R)  144, 145, 152, 156, 164, 194–195
PEG see polyethylene glycol
peptide group  71, 72
pgChem:Tigress  38
pharmacophores  289–291
3D  281, 283, 291–294
database creation  291–292
definition  283, 313, 350
design  333–337
hit-lists  292, 305
ligand-based models  298–301
model creation  288, 291, 294–301
model editing  301–303
properties  106
search  337–341
structure-based models  289, 294–298
validation  290–291
virtual screening  291–294, 301–307
see also pharmacophorical profiling
pharmacophorical profiling  379–391
aim and context  380
binding site comparison  388, 389
description of the query  381
description of the searched dataset  381
example case  380–381
exercises  387–390
input data  385–386
ligand-centric method  379, 380
methods  381–386
phenylalanine  370, 390
phenyl xanthine  181
Phi-Psi angles  320
phosphodiesterase (PDE5)  119
Pipeline Pilot  395
pi-pi interactions  328
platelets  314
PLIF see protein-ligand interaction fingerprints
PLS see Partial Least Squares
PMML (Predictive Model Markup Language)  26
poison gas  356
polarizability  132
poloxamines  9
polyethylene glycol (PEG)  9
polypropylene glycol (PPG)  9
positive ionizable features  284
posterior probability  225
PPG see polypropylene glycol
PRC AUC  178, 187
PRC see precision-recall curves
precision  177, 178, 225
precision-recall curves (PRC)  178
principal component analysis (PCA)  94–95
visualization of high-dimensional data  113–115
principal components  94–95
probability  179
protein-centric method  379–380
Protein Data Bank (PDB)  175, 193, 313, 314, 359, 381
protein-ligand binding  313, 325–328
protein-ligand docking  355–375
example case  356
exercises  362–372
input data  360–361
ligand preparation  359
methods  356–360
parameters  360
preparation of AChE  366–367
protein preparation  359–360
of tacrine  363–365
protein-ligand interaction fingerprints, (PLIF)  341, 346–347
protein-protein interfaces  281
proteins  281, 286
alternates  322
analysis and preparation  316–330, 359–360
AName  323
binding sites  389
cavities  384, 389
charge  325
detection of cavities  328–330
geometry  320
HCount  322
input for docking 359–360
quality analysis 320
structure preparation 321–325
termini 322
virtual screening 313–350
see also protein-ligand binding; protein-ligand docking; protein-ligand interaction fingerprints, (PLIF); protein-protein interfaces
proteolysis 314
prothrombin 314
protonation 89, 325, 356
pruning sets 258, 261
pseudo-random number generator 77
psychoses 380
PubMed 23, 86, 87, 175
PubMed BioAssay 175, 193, 223
PubMed 292
pulmonary embolism 314
pydot 223, 232
pyridine 71
pyridoxine 68
pyrimidine carboxamide 198, 205
pyrrolidine 326
pyrimidin-2-one 22
Python,
  atom environment fingerprints 441, 443
  CANON algorithm 430–431
  canonical SMILES 420, 423
  compound classification 223
  compound libraries 85–87, 91
data structures 415–416
hash function 443
initial invariants 427
search strategy 405
similarity metrics 403
substructure searching 432, 435–436
Ullmann algorithm 437, 438
virtual screening 395–396, 414–415

q
QSAR see quantitative structure-activity relationship
QSAR/QSPR modeling 209, 222
QSPR modeling 135
  see also ISIDA_QSPR
quality assessment 237–238
quantitative structure-activity relationship (QSAR) 129
cross-validation 163–164
machine-learning methods 209
quantitative structure-property relationship (QSPR) 129
“quorum control” 143, 153

r
R 105, 107–117
Radial Distribution Function code (RDF descriptors) 130–132
Ramachandran plot 320
Rand Index 112, 113, 116
Random Forest classification algorithm 210, 217
Random Forest method (RF) 264, 267–269
Random Subspace method (RS) 264–267, 269
Random Tree method 264
RDBMS see relational database management system
RDF descriptors see Radial Distribution Function code
RDKit 85, 86, 395, 416
  AddAtom 88
  atom methods 418–419
  bond methods 419
  Compute2DCoords 87
  creation of molecule objects 417
descriptor calculation 90
generation of fingerprints 399
HierarchicalClusterPicker 97
Hill notation for molecules 419–420
immutability 88
MaxMinPicker 97–98
molecule methods 418
MurckoScaffold module 98–99
ReactionFromSmarts 99
RemoveBond 88, 89
RunReactants 99
RWMol 88
SaltRemover 89
SDMolSupplier 87
SMIRKS strings 99
RD-RF see Regression by Discretization based on Random Forest
REC see regression error characteristic curve
recall 177, 225
see also precision-recall curves (PRC)
receiver operating characteristic curve (ROC) 178, 237, 246, 290, 292–294
virtual screening 407–408, 410–412
re-docking 341–345, 355–356
of donepezil 370
of tacrine 362
refTversky 384, 390
Regression by Discretization based on Random Forest (RD-RF) 210, 217
regression equations 142
regression error characteristic curve (REC) 197, 198, 202
regression models 193–208
bagging and boosting 249–255
step-by-step instructions 197–206
theoretical background 194–198
Regression Tree (M5P) 210, 217, 271, 274–275
relational database management system (RDBMS) 37–65
chemically aware interface 38
step-by-step instructions 41–65
theoretical background 37–41
relative absolute error 221
relative mean squared error 220
RF see Random Forest method
R-groups 67, 68, 71, 73, 98–99
ethoxy 71
ethyl 71
methyl 71
Ridge Regression, (RR) 194, 197–198, 202, 207
RIPPER machine learning method 259
Ripper k algorithm 258
RMSD see root-mean-square deviation
RMSE see root-mean-square error
ROC see receiver operating characteristic curve
ROC AUC 178, 187, 190, 246, 267
ROCKS 381, 384, 387
output files 387
profiling 388
root-mean-square deviation (RMSD) 363–365, 367, 370
see also root-mean-square error (RMSE),
root-mean-square error (RMSE) 144, 145, 187, 194, 196, 254
estimation 148–152
origins 194
random subspace 266–267
stacking 277
see also root-mean-square deviation (RMSD)
RR see Ridge Regression
RS see Random Subspace method
RStudio 107
rule growing 258
rule pruning 258
S
salts, removing 89
Sanderson electronegativity 132
SBCADD see structure-based computer-aided drug design
scatterplots 93–95, 196
schizophrenia 380
Schrödinger 38, 315
scikit-learn 85, 86
compound classification 223–239
principal components 95
score 178
score plots 95
scoring functions 355
SDF files 87, 136, 137
editing 139
MDL 78, 131
SD files 87, 139, 144, 153, 155, 399–400, 417
search strategies 337–341, 397, 405, 432, 435–436
selective drugs 379
sensitivity 177
Sequential Minimal Optimization (SMO) 180, 181, 187, 201, 210, 215
serine proteases 314
serotonin 5-HT receptor ligands 105, 396, 412
shape comparison 379–391
Shaper 381, 384, 389–390
shrinkage parameter 250, 253, 254
silicates 22
silica 22
similarity function 397, 404
similarity-property principle 396
similarity searching 396
slack variables 227
small organic molecules 3, 21, 325
SMARTS 89, 432
SMILES 3, 4
notation 76, 89, 416, 420–422
strings 75–80, 422–425
SMIRKS 99
SMO see Sequential Minimal Optimization (SMO)
Soergel distance 106, 108
split criterion 232–233
split strategy 233
stacking 271
algorithm 272–273
step-by-step instructions 273–278
theoretical background 271–272
standard deviation 150, 152, 156, 196
standardization 46, 88, 91, 94, 332, 359
stearths 9
stereochemistry 4, 90
steric clashes 320
structure-based computer-aided drug design (SBCADD) 314, 315, 320
structure-property relationships 75
Student’s criterion 143
Student’s t-tests 219
subgraph isomorphism problem 416–417
substructures 77
searching 432–441
sulfur 287
supervised machine learning methodology 224
compound classification 223, 226–227, 234–236, 238
e-insensitive (e-SVM) 194
Support Vector Regression (SVR) 210, 215
SVD Singular Value Decomposition
SVM see Support Vector Machine
SVR see Support Vector Regression
Sybyl 359
tables, for storing information 37
tacrine 356, 359, 362–366, 372
tacrine-pyridone 366
inhibitor 367–370, 372
Tanimoto coefficient (Tc) 78, 105, 106, 108–110, 115, 150, 397, 402–404
TanimotoCombo score 388
Tanimoto score 381
target fishing 281
taste description 4, 10–12, 15, 28–29, 31, 34, 37, 38
Tc see Tanimoto coefficient
tetrahedral centres 76
tetrahydroimidazobenzodiazepine (TIBO) derivatives 136
thiols 287
thiopene 285
active 314
as a drug target 314
Ramachandran plot of 320
three-dimensional structure 315
thrombin/inhibitor PDB complex 320
thrombocytopenia 314
thrombosis 314
TIBO see tetrahydroimidazobenzodiazepine (TIBO) derivatives
titanium oxide 22
TLSER descriptors 209
toluene 71
Tompson’s rule 153
topological paths 140–142, 154
Torpedo californica 356
toxicity 130
TPR see True Positive Rate
triazolopyrimidine derivatives 200
trideseths 9
Tripos 315
True Negative Rate (TNR) 177
True Negatives (TNi) 176
True Positive Rate (TPR) 177, 178, 293
True Positives (TPi) 176, 407
tryptophane 390
tyrosine 390
tyrosine kinase inhibitors 105

U
Ullman algorithm 433, 436–441
Uniprot 314
unit variance scaling 113
USPTO 68

V
vanadium pentoxide 22
Vancomycine 4
van der Waals radii 288
van der Waals surface 130
van der Waals volume 132
variable selection bias 163
vertex invariant 426
virtual screening (VS),
benchmarking performance 406–414
completed program 405–406
docking approach 341–348, 350
introduction to 396
methods 313–314
pharmacophore approach 291–294,
313–314, 330–341
pillars of 397
proteins 313–350
retrospective 330–348
search strategy 397, 404–405
similarity metrics 402–404
visualization,
of decision tree 232, 234
of descriptor distributions 92–94
of fingerprints 113, 115

W
Ward’s criterion 107
weight vector 226–227
WEKA,
  bagging and boosting 243–245, 249–255
  classification models 175, 180–181,
  185, 243–245
  cross-validation 163, 165–166, 168,
  171–172
  data visualisation 119, 122
  descriptor generation 131, 134
  Experimenter mode 212
  Ibk method 163, 180, 216, 271
  interpretable rules method 257–261
  machine learning methods 209–221
  Random Forest method 263–267
  regression models 193, 197, 200, 204,
  249–255
  stacking 271–277
Wermuth, Camille 283, 313
widgets 60
Wiener index 130, 131
wrapper method 165, 168

X
xanthine 205
derivatives 198
XLogP descriptor 131
x-ray crystallography 315, 334

Z
Zero Regression (ZeroR) 168, 209–210
ZINC12 396
ZINC database 75, 332, 346, 406
z-scores 91–92, 321