Ab initio predictions, lead optimization, overview, 149–154
Absorption, distribution, metabolism and excretion (ADME):
Chemical Property Structure Planning System (BICEPS), 295–297
chemical array analysis, 179–181
global model development and applications:
ligand-based models, 247–248
metabolic lability, 248–262
data set characteristics, 251
descriptor sets, 251–256
DGAT1 inhibitor optimization, 256–262
machine learning and molecular descriptors, 249–251
structure-based models, 246–247
naïve Bayesian models, target identification, 137
Accuracy of classification, rough set theory, 56–61
Acetylcholinesterase (ACT), virtual screening, 122–123
Activity cliffs/landscapes:
case study, 308–309
chemical arrays, 185–186
lead optimization, 156–157
molecular similarity applications, 383–384
structure-activity relationships:
antitarget activity hotspots, 228–230
classical chemical studies, 214–216
outlier investigations, 209–210
pharmaceutical applications, 223
quantification, 210–211
Activity prediction models:
naïve Bayesian models, phenotype/target comparison, 141–143
three-dimensional QSAR, 45–47
ActWiki, chemical arrays, 202
Adenosine di/triphosphate (ADP/ATP), ecto-5’-nucleotidase virtual screening, 125–128
ADME assays, regression models, 6–7
Adverse drug reactions (ADRs):
naïve Bayes models, 137
predictive secondary pharmacology, 274–276
Alignment rules:
molecular similarity, 348–350
three-dimensional-ligand-based models, 225–227
Aliphatic side chain substitutions, Topliss Tree, 150–152
Ames QSAR model, toxicity warning systems, 276–278
Aminergic GPCR inhibitory activity, partial least squares analysis, 86–87
Aminomethylcyclohexane (AMC), volumetric/surface-density functions, molecular similarity models, 41–47
2-Aminothiazoles, metabolic lability model, DGAT1 inhibitors, 256–262
Annotation:
chemical arrays, 194–195
controlled substances, 314–315
Antitarget activity:
in silico screening, ligand-based ADMET models, 247–248
structure-activity relationships, 227–232
activity hotspot identification and application, 228–230
hERG and CYP3A4 inhibition, 230–231
lead optimization integration, 232
transfer mechanisms, 227–228
Apostle approach:
  chemoinformatics database rollout, 338
Mobius CIDB system, 329
Approximation set, rough set theory, 55–56
AQUASOL database, predictive modeling, 3–6
Aqueous solubility, predictive model comparisons, 16–17
Area under the curve (AUC) metrics, fingerprint method comparisons, 104–110
Area under the receiver operating characteristic curve (ROCAUC), naïve Bayesian models:
  chemist’s preferences, 144–145
  potency binning, 140–141
Aromatic ring substituents, Craig plot, 150–151
Association rules, rough set theory, 63–64
Asymmetric similarity:
  mathematical description, 350
  molecular similarity analysis, 366–370
  two-dimensional searching, 368–370
Atomic colorings:
  L-shaped partial least squares analysis:
    molecular modeling, 91–93
    regression coefficient matrix, 90–91
AurSCOPE database, antitarget activity analysis, 229–230
Automated annotation workflow, controlled substances, 314–315
Automatic chemotype detector, chemical arrays, 195–198
AZOrange software, safety modeling, 283–285
Bayesian analysis:
  L-shaped partial least squares analysis, atomic colorings, regression coefficient matrix, 90–91
toxicity studies, 279
Bayesian weights, naïve Bayesian models, class comparisons, 143
BCUT descriptors, molecular similarity, 356
Beacon Projects, 328
Benchmark studies, virtual screening techniques, 115–116
Benzamidine, volumetric/surface-density functions, molecular similarity models, 41–47
Binary fingerprints, molecular similarity, set-based representation, 353
Bioisosterism:
  molecular fingerprinting techniques, 99–100
  rescaffolding, 223–227
Biological activity:
  naïve Bayesian models, 139–140
toxicity studies, 279–283
Biological response surface, activity cliffs, outlier investigations, 210
BioProfile system, 302–307
BioSim system, toxicity analysis, 279–283
Boehringer Ingelheim Comprehensive Library of Accessible Innovative Molecules (BICLAIM) database:
  development of, 293
  search strategies, 309–314
Boehringer Ingelheim Mining and Exploration of Screening Hit (BIMESH), 299–301
Boehringer Ingelheim Split Substructure Search (BISCUBE), BICLAIM-space searches, 311–314
Boehringer Ingelheim Chemical Property Structure Planning System (BICEPS), 295–297
BIMESH HTS data analysis, 300–301
BioProfile system, 302–307
Chemoinformatics Database (CIDB), 293–294
controlled substances, automatic annotation, 314–315
Database of Virtual Combinatorial Libraries, 297–299
  search applications, 309–314
Boltzmann-Enhanced Discrimination of ROC (BEDROC) metric, molecular similarity validation, 377–378
Box counting procedures, molecular similarity, chemical space representation, 380–383
Boxplots, datasets, 4–6
Build vs. buy, chemoinformatics database systems, 339–340
Bump hunting methods, multiparameter lead optimization, 168–169
Candidate marker genes, drug-induced gene expression, phospholipidosis:
  identification, 66–68
  ranking, 75–77
Cardiac liability risk reduction, QSAR safety filter, 271–272
Cardinality, set-based similarity functions, molecular similarity, 357–359
Cavity-based alignment, aminergic G protein-coupled receptors, 87–88
Cell-based chemical space representation, molecular similarity, 378–383
Central compound database (CDB) system, 292–293
ChEMBL target families, naïve Bayesian models, 140–141
Chemical arrays, lead optimization:
  annotation, 194–195
archive information extraction, 194–201
automatic chemotype detector, 195–198
chemistry and property space coverage, 182–186
data analysis techniques, 192–194
seed compound detection, 196, 198–201
self-avoiding random walk, 191–192
temporal analysis, 186–191
Chemical graph-based representations, molecular similarity, 353–354
applications, 378–383
similarity functions, 361–363
Chemically advanced template search (CATS): metabolite lability model, 250–251
rescaffolding, 224–225
Chemical similarity: metabolite lability model, 253–256
molecular fingerprinting techniques, 98–100
Chemical space representation:
activity cliffs, heatmaps, 215–216
machine learning, 138
molecular similarity, 350–356
graph-based representations, 353–534
set-based representation, 352–353
vector-and function-based representations, 354–356
multiparameter optimization, hit to candidate, 174–175
Chemical structure analysis. See also Molecular similarity analysis (MSA)
toxicity studies, 272–278
Chemiluminescent nitrogen detection (CLND), PubChem, 4–6
ChemistryConnect warehouse, safety modeling, predictive secondary pharmacology, 275–276
Chemistry space plots, chemical arrays, 182–186
ChemLink CIDB, 325–327
Chemogenomics, quantitative structure-activity relationship, 86
Chemogenomics studies, partial least squares method, 85–86
aminergic GPCR inhibitory activity data, 86–87
atomic colorings:
molecular modeling, 91–93
regression coefficient matrix, 90–91
LPLS analysis, 91
LPLS ligand and protein descriptors, 87–88
L-shaped PLS architecture, 89–90
Chemoinformatics databases (CIDs):
apostle-based rollout systems, 338
Beacon Projects, 328
bottom-up vs. top-down, 338
build vs. buy vs. open source, 339–340
ChemLink, 325–327
continuity, 337
Cousin system, 323–325
data sources and characteristics, 336
Mobius CIDB system, 328–335
off-the-shelf software, 338
quality software requirements, 336
RGate 2003+ CIDB, 327
support/maintenance requirements, 339
training requirements, 339
user needs assessment, 336
Chemometrics, naïve Bayesian models:
data types and quality, 139–141
enriched features, mining and interpretation, 143–145
molecular representations and machine learning, 138
target and phenotype comparisons, 141–142
virtual screening, 134–137
Circular fingerprints, two-dimensional analysis, 100–101, 110–111
Classification models, predictive applications, 6–7
Classification of accuracy, rough set theory, 56–57
Cluster analysis, 301
CODD system, chemical arrays, 202
Cognitive science, molecular similarity and, 347–350
Combination of evidence, toxicity studies, 278–279
Combinatorial chemistry, BioProfile system, 302–307
CoMFA model:
activity cliff quantification, 210–211
rescaffolding, 225
visualization in SAR analysis, 217–222
Committee models, metabolic lability, 252–256
Compass 3D-QSAR approach, molecular similarity models, 40–47
Competing objectives, lead optimization, 171–173
Compound structure:
metabolic lability model, 253–256
naïve Bayesian models, 139
class comparisons, Bayesian weights, 143
enriched feature interpretation, 143–145
Computable similarity function, molecular similarity, 350
Computational techniques:
drug design, 35–47
naïve Bayesian models, off-target predictions, 137
toxicity analysis, current research trends, 267–269
virtual screening, 115–116
Computational Theory of Mind, 389
Condition attributes:
reducts, 61–63
rough-set theory, 52
Confidence intervals, predictive model
comparisons, 15–17
Confusion matrix, metabolic lability model, 252–256
Consensus activity cliffs, 216
Coordinate-based chemical spaces, molecular
similarity, 378–383
Correlation coefficient, predictive model
comparisons, 15–17
Correlation methods, molecular similarity
measurements, 375–376
Cousin CIDB system, 323–326
Craig plot, lead optimization, 150–154
C-reducts, 61–63
Cubist software, metabolic lability model, 249–251
Customer interface, safety profiles, 283–285
Cyclopentyl ethers, metabolic lability model,
DGAT1 inhibitors, 262
CYP3A4 inhibitor:
antitarget activity, 228–230
atom colorings, Bayesian analysis, 90–91
metabolic lability model, 248–256
DGAT1 inhibitors, 257–262
structure-activity relationships, inhibition
mechanisms, 230–231
structure-based ADMET models, 246–247
Cytochrome P450s (CYPs):
metabolic lability model, 248–256
structure-based ADMET models, 247
Cytohesin inhibitors, virtual screening models,
124–125
Dabigatran, BICLAIM-space searches, 312–314
Data-driven modeling, safety analysis, 269–283
Data fusion methods, molecular similarity
analysis, 371–373
Data mining, privileged substructures, 209
Dataset characteristics:
chemoinformatics databases, 336
drug-induced gene expression,
phospholipidosis, 68–71
lead optimization, retrospective analyses,
171–173
matched molecular pairs (MMPs), 212–214
metabolic lability model, 251
Mobius CIDB system, 330–332
naïve Bayesian models, 139–141
structural-activity relationships, pharmaceutical
applications, 222–223
Daylight similarity:
molecular fingerprinting techniques, 98–100
molecular fingerprint method comparisons,
103–104
Decision analysis techniques, lead optimization,
175–176
Decision attributes, rough-set theory, 52
Decision table (DT):
drug-induced gene expression,
phospholipidosis, 68, 71–72
genomics research, 51–52
rough set theory, 53–56
Decision trees, metabolic lability model, 249–251
DEGAS system, chemical arrays, 202
De Morgan’s Laws, set-based similarity functions,
molecular similarity, 357–359
Dempster-Shafer theory (DST), toxicity studies, 279
De novo drug design:
nâä¹î Bayesian model limitations and, 146–147
rescaffolding, chemically advanced template
search, 224–225
safety QSAR modeling, 270–271
Desirability index, multiparameter lead
optimization, 160, 162–164
Diacylglycerol-acyl-transferase 1 (DGAT-1)
inhibitors:
metabolic lability model, 256–262
three-dimensional ligand-based relationships,
226–227
Digital PDP-11 graphics system, 323
Dimensionality reduction, visualization in SAR,
216–222
Discernibility classes, rough set theory, 59–61
Discretization analysis, drug-induced gene
expression, phospholipidosis, 68–71
Dissimilarity plots:
chemical arrays, 182–186
molecular similarity and, 348–350
Distance measurements, molecular similarity, 350
Distance-to-model methods, safety QSAR
modeling, 270–271
Distill program:
molecular scaffolds, 208
visualization in SAR and, 216–222
Downregulated scores, drug-induced gene
expression, phospholipidosis, 68–71
DRAGON descriptors, 249–256
D-reducts, 61–63
drug-induced gene expression,
phospholipidosis, 68, 71–72
Drug design:
incrementalism and serendipity, 33–35
physical processes and computational methods,
35–47
molecular similarity, 39–45
protein structure-based methods, 37–39
three-dimensional QSAR, 45–47
toxicity warning systems, 276–278
virtual screening, 121–123
Drug-induced gene expression, phospholipidosis, 65–77
data set characteristics, 68–71
D-reduct determination, 68, 71–72

Ecto-5’-nucleotidase, virtual screening, 125–128
Edge effects, classification models, 6–7
Electronic lab notebook (eLNB), chemical array annotation, 194–195
seed compounds, 199–201
Electronic laboratory network (ELN), controlled substances annotation, 315
Electrostatic similarity, molecular similarity models, 44–45
Endpoint analysis, in vitro studies, safety QSAR modeling, 269–270
EON system, three-dimensional ligand-based relationships, 226–227
Exclusive-or (XOR) problem, drug design and, 35–47
Expectation values, Tanimoto similarity, 102
Extended-connectivity fingerprints of depth 4 (ECFP_4), naïve Bayesian models, 133–134
virtual screening and, 134–137
Extended-connectivity fingerprints of depth 6 (ECFP_6):
aminer GPCR inhibitory activity, 86–87
L-shaped partial least squares analysis, atomic colorings, regression coefficient matrix, 91
Extreme programming, safety modeling, 283–285

Factorial design, lead optimization, 151–152
Factor Xa inhibitors:
model diagnosis plots, 210–211
R-group plots, 214–215
visualization in SAR analysis of, 217–222
FCFP4 fingerprints, metabolic lability model, 254–256
Feature-based fingerprints:
cognitive science and, 348–350
molecular similarity applications, 98–100
naïve Bayesian models, enrichment and interpretation, 143–145
Feature trees, rescaffolding, 224–225
Filter techniques, multiparameter lead optimization, 159–160

First-in-class inhibitors, virtual screening, 125–128
Fisher’s $z'$ distribution, predictive model comparisons, 15–17
Fixed-length fingerprints, molecular similarity, set-based representation, 352–353
Fraction of sp3 carbons, multiparameter lead optimization, 158–159
Fragility limitations, molecular fingerprinting techniques, 98–100
circular fingerprints, 100–101
Frequent hitter analysis, 304–307
Fuzzy relations, molecular similarity, 350
chemical space representation, 380–383
set-based representation, 353
set-based similarity functions, 357–361

Gaussian reward functions:
molecular similarity models, 40–45
structure-activity relationships, three-dimensional ligand-based relationships, 225–227
Genetic algorithms:
lead optimization, 172–173
metabolic lability model, 249–256
Genetox Warning System (GWS), 276–278
GOLD molecular docking technique, protein-ligand binding, 38
G protein-coupled receptor (GPCR) inhibitory data:
L-shaped partial least squares analysis:
amolecular modeling, 92–93
regression coefficient matrix, 90–91
ligand and protein descriptors, 87–88
naïve Bayesian models, 145–146
quantitative structure-activity relationship, 86
aminergic GPCR inhibitory activity data, 86–87
safety modeling, predictive secondary pharmacology, 275–276
Graph-based algorithms, maximum common substructure, 208
Graphic-user interface (GUI), chemoinformatics databases, 326
GRASP visualization program, molecular fingerprinting, 111
Group fusion, molecular similarity analysis, 372

H2 antagonists, multiobjective lead optimization, 154–157
Hansch values, lead optimization, 151–154
Heatmap representation, activity cliff analysis, 215–216
HierS scaffold clustering system, molecular scaffolds, 208

High-throughput screening (HTS):
activity cliffs, outlier investigations, 209–210
lead optimization, retrospective analyses, 171–173
molecular fingerprinting techniques, 99–100
molecular scaffolds, 208
multiobjective lead optimization, 155–157

Histamine 3 receptors, safety modeling, predictive secondary pharmacology, 275–276

Historical data, multiparameter lead optimization, 168–169

Hit identification. See also Lead optimization
multiparameter optimization, hit to candidate, 174–175
virtual screening:
applications, 116–117, 123–128
first-in-class inhibitor, 124–125
multifunctional protein inhibitors, 124–125
definition and classification, 113–114
future research issues, 128
high-throughput screening vs., 119–121
molecular drug targets, 121–123
performance evaluation, 117–119

Human ether-à-go-go related gene (hERG):
chemical arrays, 202
chemotype detection, 196, 198
off-target activity models, 1, 137
oxadiazole topology and, 213
regression models, 7
scoring profile, 165
structure–activity relationships:
antitarget activity hotspots, 228–230
inhibition mechanisms, 230–231
temporal analysis, 191
three-dimensional QSAR model, 246–247
toxicity analysis:
cardiac liability risk reduction, 271–272
current research trends, 268–269

Human hepatoma HepG2 cells, drug-induced gene expression, phospholipidosis, 65–77

Hungarian algorithm, molecular fingerprinting, 111

Huuskonen dataset:
predictive modeling and, 3–6
applicability, 13–15
random forest model construction, 9–11

Hydrogen bond donors and acceptors multiparameter lead optimization, 158–159
desirability functions, 163–164
Hydrophobic ligands, protein-ligand binding, 37–39

Icaris system, chemoinformatics databases, 328

IF-THEN rules:
drug-induced gene expression,
phospholipidosis, preliminary rules generation, 72–75

Imprecision, similarity analysis, 98–100
Incremental modification, drug design, 34–35

Indiscernibility classes:
drug-induced gene expression,
phospholipidosis, D-reducts, 71–72

Innovative Medicines Initiative (IMI) eTOX project, toxicity analysis, 282–283

In silico predictions:
ADMET global model development and applications:
future research issues, 262
ligand-based models, 247–248
metabolic lability, 248–262
dataset characteristics, 251
descriptor set results, 251–256
DGAT1 inhibitor optimization, 256–262
machine learning and molecular descriptors, 249–251
structure-based models, 246–247
multiparameter optimization, hit to candidate, 174–175

safety modeling, 268–269
preclinical data, 282–283
predictive secondary pharmacology, 275–276
three-dimensional ligand-based relationships, 226–227

Integrated chemoinformatics systems:
postulate-based rollout systems, 338
Beacon Projects, 328
bottom-up vs. top-down, 338
build vs. buy vs. open source, 339–340
ChemLink, 325–327
Cousin system, 323–325
data sources and characteristics, 336
Mobius CIDB system, 328–335
off-the-shelf software, 338
quality software requirements, 336
RGate 2003+ CIDB, 327
succession planning, 339
support/maintenance requirements, 339
training requirements, 339
user needs assessment, 336

Interquartile range (IQR), datasets, 4–6
Intersection sets, set-based similarity functions, molecular similarity, 357–359

In vitro analysis:
ADMET global model development and applications, 245–248
multiparameter optimization, hit to candidate, 174–175
naïve Bayesian models, 143–145
toxicity studies:
  biological data, 281–283
  chemical structure links, 269–272
  preclinical data, 282–283
  safety QSAR endpoint modeling, 269–270
  warning systems, 276–278

*In vivo* analysis:
  multiparameter optimization, hit to candidate, 174–175
toxicity studies:
  *in vitro* profile data, 281–283
  warning systems, 277–278
ISIS components, chemoinformatics databases, 326, 328
Isostere generation, molecular fingerprinting techniques:
  basic principles, 99–100
  WABE programming, 101–102
IUPAC International Chemical Identifier (InChI), 139
Kendall’s tau, predictive model performance evaluation, 7–8
molecular similarity, 14–15
Kernel loadings matrices, L-shaped partial least squares, 90
*k*-nearest neighbor (kNN) QSAR:
  metabolic lability model, 249–256
  Similarity Principle, 207
KNIME workflow tool:
  BioProfile system 304–307
  Boehringer Ingelheim Mining and Exploration of Screening Hit (BIMESH) system, 300–301
  Boehringer Ingelheim project data marts, 297
  Boehringer Ingelheim workflow system, 294–295
Kohonen networks, visualization in SAR and, 219–222

Lead-hopping, molecular fingerprint method comparisons, 103–104

Lead optimization:
  antitarget activity, 232
  chemical arrays:
    annotation, 194–195
    archive information extraction, 194–201
    automatic chemotype detector, 195–198
    chemistry and property space coverage, 182–186
    data analysis techniques, 192–194
  seed compound detection, 196, 198–201
  self-avoiding random walk, 191–192
  temporal analysis, 186–191
chemoinformatics:
  overview, 149–154
multiobjective methods, 158–169
  basic rules, 158–159
  desirability functions, 160–164
  filters, 159–160
  probabilistic scoring, 164–165
  property profile, 165–169
  multiparameter optimization, hit to candidate, 174
retrospective analyses, 169–173
structure–activity relationships (SARs):
  activity cliffs:
    exploration, 214–216
    outlier investigation, 209–210
    quantification, 210–211
    antitarget activity, 227–232
    activity hotspot identification and application, 228–230
    hERG and CYP3A4 inhibition, 230–231
    transfer mechanisms, 227–228
    matched molecular pairs, 211–214
    molecular scaffolds, 207–208
    overview, 205–206
pharmaceutical industry applications, 222–223
  privileged substructures, 208–209
  rescaffolding, 223–227
    three-dimensional-ligand-based approaches, 225–227
    three-dimensional-protein-based approaches, 227
    two-dimensional approaches, 224–225
    similarity principle, 207
    visualization support, 216–222
LeadScope program, visualization in SAR analysis, 216–222
Learned ideal distance, molecular similarity models, 40–45
Lexicographic fingerprints, molecular similarity applications, 98–100
LHASA project, 323
Ligand-based ADMET models, properties of, 247–248
Ligand-based virtual screening (LBVS):
  applications, 123–128
  correlation methods, molecular similarity and, 375–376
  cytohesin inhibitors, 124–125
  definition and classification, 113–114
  performance evaluations, 117–119
  practical applications, 116–117
Ligand descriptors, L-shaped partial least squares analysis, 87–88
Ligand-efficiency (LE) analysis, 308–309
Linear cascade approach, temporal analysis, lead optimization, 187–191
LINGO fingerprint method:
   molecular similarity applications, 98–100
two-dimensional analysis, 101
Linguistic rules, rough set theory, 55–56
Lipinski’s Rule of Five:
   ADMET global model development and applications, 246–248
   multiparameter lead optimization, 158–159
LiSARD interactive graphics, visualization in SAR and, 219–222
Local neighborhood plots, visualization in SAR and, 219–222
Log P:
   multiparameter lead optimization, 158–159
temporal analysis, lead optimization, 186–191
Log ratios, naïve Bayesian models, 133–134
Log S, random forest model construction, 9–11
L-shaped partial least squares (LPLS):
   chemogenomics studies:
      aminergic GPCR inhibitory activity data, 86–87
      atomic colorings:
         molecular modeling, 91–93
         regression coefficient matrix, 90–91
   LPLS analysis, 91
   LPLS ligand and protein descriptors, 87–88
   L-shaped PLS architecture, 89–90
MACCS keys fingerprint:
   metabolic lability model, 250–251
   molecular similarity applications, 98–100
   chemical space representation, 380–383
   statistical independence, 374–375
two-dimensional analysis, 100–101
Machine learning:
   drug design and, 36–47
   metabolic lability model, 249–256
   naïve Bayesian models, comparisons with, 138–139
   safety QSAR modeling, 270–271
   Magic methyl phenomenon, lead optimization, 156–157
   Maintenance systems, chemoinformatics database systems, 339
   Mantel statistic, molecular similarity techniques, 376–377
   MARS program:
      antitarget activity analysis, 230
Molecular Data Explorer program, structure similarity maps, 219–222

Molecular descriptors:
- ligand-based ADMET models, 247–248
- metabolic lability model, 249–256
- molecular similarity, 350–356
- graph-based representations, 353–534
- set-based representation, 352–353
- vector and function-based representations, 354–356
- naïve Bayesian models, 138–139
- safety QSAR modeling, 270–271
- structure-activity relationships, activity cliffs, outlier investigations, 210

Molecular docking techniques:
- protein-ligand-binding and, 37–39
- structure-based ADMET models, 247

Molecular equivalence number structural classification system (Meqnum), molecular scaffolds, 208

Molecular fingerprinting:
- chemical arrays, 182–186
- CYP enzymes, 247
- metabolic lability model, 253–256
- molecular similarity:
  - set-based representation, 352–353
  - set-based similarity functions, 357–361
  - weighted representations, 356–357
- naïve Bayesian models, 138–139
- safety modeling, predictive secondary pharmacology, 275–276
- stability analysis:
  - isostere generation, WABE program, 101–102
  - Tanimoto similarity, 102
  - two-dimensional methods, 100–101
  - toxicity studies, 281–283

Molecular scaffolds, structure-activity relationships, 207–208

Molecular similarity analysis (MSA):
- activity landscapes and cliffs, 383–384
- asymmetric similarity, 366–368
- chemical space representation, 350–356, 378–383
- graph-based representations, 353–534
- set-based representation, 352–353
- vector and function-based representations, 354–356
- cognitive aspects, 347–350
- comparison of measurement methods, 375–377
- data fusion and consensus methods, 371–375
- functions/coefficients, 357–365
- chemical graph-based similarity functions, 361–363
- set-based similarity functions, 357–361
- vector/function-based similarity functions, 363–365
- metabolic lability model, 253–256
- molecular fingerprinting:
  - isostere generation, WABE program, 101–102
  - Tanimoto similarity, 102
- molecular representation, 350–356
- graph-based representations, 353–354
- set-based representation, 352–353
- vector and function-based representations, 354–356
- SAR and QSAR techniques, 97–100
- statistical independence, 374–375
- structure-activity similarity and related maps, 384–387
- two-dimensional asymmetric similarity searching, 368–370
- validation of measurements, 377–378
- weighted representations, 356–357

Molecular weight, multi-parameter lead optimization, 158–159
desirability functions, 163–164

MolPrint 2D fingerprints, chemical arrays, 182–186

MPP12 inhibitors, lead optimization, 172–174

mRNA scores, drug-induced gene expression, phospholipidosis, 68–71, 77–79

Muchmore-Martin lead-hop sets, molecular fingerprint method comparisons, 103–104

Multiclass naïve Bayesian models, potency binning, 140–141

Multidimensional scaling, visualization in SAR analysis, 219–222

Multifunctional protein inhibitors, virtual screening models, 124–125

Multi-fusion similarity, 373

Multiobjective methods:
- lead optimization, process overview, 154–157
- temporal analysis, lead optimization, 186–191

Multiparameter optimization (MPO), lead optimization, 158–169
- basic rules, 158–159
- desirability functions, 160, 162–164
- filters, 159–161
- hit to candidate process, 174–175
- probabilistic scoring, 164–167
- property profile, 165–169

Multiple feature tree model (MTree), rescaffolding, 224–225

Multiplicative methods, lead optimization, desirability functions, 163–164
Multiset fingerprints, molecular similarity, set-based representation, 352–353
Multitargeted assays, 132–134, 216
Muscarinic antagonists, molecular similarity models, 44–45

Naïve Bayesian models (NBMs):
chemometric applications:
data types and quality, 139–141
enriched features, mining and interpretation, 143–145
target and phenotype comparisons, 141–142
virtual screening, 134–137
safety modeling, predictive secondary pharmacology, 275–276

Nearest neighbor techniques:
molecular similarity, chemical space representation, 378–383
safety QSAR modeling, 270–271

Neighborhood plots, activity cliff quantification, 210–211

Network-like similarity graphs (NSGs):
activity cliff analysis, 215–216
virtual screening, 121–128
visualization in SAR, 221–222

NIPALS algorithm, L-shaped partial least squares, 89–90

Nondeterministic rules, rough-set theory, 64

Nonlinear structure-activity relations:
multiobjective lead optimization, H2-antagonists, 154–157
visualization techniques, 219–222
Norepinephrine transporter inhibitors, structure-activity similarity (SAS) maps, 385–387

Off-target candidates:
drug design, 35
naïve Bayesian models, 137
predictive secondary pharmacology, 274–276

Off-the-shelf software, chemoinformatics databases, 338

Open source platforms, chemoinformatics database systems, 339–340

Over-robustness, molecular fingerprinting techniques, 99–100
Oxadiazole derivatives, matched molecular pairs (MMPs), 212–214
Oxidative metabolic clearance, metabolic lability model, 248–256

PARASURF system, metabolic lability model, 250–252

Pareto optimization:
  multiparameter lead optimization, 166, 168–170
temporal analysis, lead optimization, 186–191

Partial least squares (PLS) method:
  activity cliff quantification, 210–211
chemogenomics studies:
  LPLS analysis, 91
  LPLS ligand and protein descriptors, 87–88
  L-shaped PLS architecture, 89–90
  visualization in SAR, 217–222

Path-based fingerprints:
  comparison with other methods, 105–110
  molecular similarity applications, 98–100
  two-dimensional analysis, 100–101
PDE5 inhibitors, protein-ligand-binding, 38–39
Pearson’s $r$:
molecular fingerprint method comparisons, 103–110
  predictive model performance evaluation, 7
  confidence intervals, 15–17
  molecular similarity, 14–15
  random forest model construction, 10–11
Perceptrons, 36–47

PFAKT system, chemical arrays, 202

Pharmacophore models, ligand-based ADMET models, 247–248

Pharmacophoric triplets, structure activity relationships, 225–227

Phenotypic screening, naïve Bayesian models:
target comparison, chemical and biological activity space, 141–143
target identification, 135–137

Phospholipidosis:
drug-induced gene expression, 65–77
dataset characteristics, 68–71
$D$-reduct determination, 68, 71–72
preliminary rule generation, 72–75
rule simplification-attribute value reduction, 75–77

Physical reality models:
drug design and, 39–45
  three-dimensional QSAR, activity prediction, 45–47

Pipeline Pilot:
  automatic chemotype detection, 195
  BICLAIM-space searches, 311–314
Boehringer Ingelheim project data marts, 297
Boehringer Ingelheim workflow system, 294–295
chemical array systems, 192
molecular fingerprinting and, 86–87
pairwise structural dissimilarity analysis, 182–183
seed detection workflow, 201
Plato system, usage facilitation and drug safety analysis, 285
Polar moieties, molecular similarity models, 44–45
Polar surface area (PSA), multiparameter lead optimization, 158–159
Polypharmacology, toxicity studies, 281–283
Potassium channel, hERG inhibition, 230–231
ligand-based ADMET models, 247–248
Potency binning, multiclass naïve Bayesian models, 140–141
Potential genotoxic impurity (PGI), current research trends, 267–269
Preclinical data, toxicity studies, 282–283
Prediction error analysis, metabolic lability model, 253–256
Predictive models:
- applicability, 12–15
- experimental error and performance, 11–12
- molecular descriptors, 8–9
- performance evaluation, 7–8
- random forest model example, 9–11
- safety QSAR modeling, 270–271
- predictive secondary pharmacology, 274–276
source code listings, 20–30
Predictive secondary pharmacology (PSP), toxicity analysis, 274–276
in vitro profile data, 281–283
Pregnane X receptor (PXR), structure-based ADMET models, 246–247
Primary target pairing, 34–35
Principal component analysis (PCA):
- molecular similarity, chemical space representation, 380–383
- visualization in SAR, 217–222
Privileged substructures, structure-activity relationships, 208–209
Probabilistic rules:
- multiparameter lead optimization, 164–167
Project coding, chemical array annotation, 194–195
Project data marts, Boehringer Ingelheim CDB, 297
Property landscapes:
- activity cliff quantification, molecular similarity, 384
- molecular similarity, 384
Property profiles:
- matched molecular pairs (MMPs), 212–214
- multiparameter lead optimization, 165, 168–169
Property space plots, chemical arrays, 182–186
Protein kinases, privileged substructures, 209
Protein-ligand binding secondary structures, 209
Protein-ligand binding interactions, 36–47
Protein pocket conformation, protein-ligand binding, 38–39
Protein structure-based methods, 37–39
Pseudoequivalences, molecular similarity, set-based representation, 353
PubChem, 4–6
- molecular similarity, 14–15
Python programming language, 2–31
Quantitative annotation, naïve Bayesian models, 139–140
Quantitative estimate of drug-likeness (QED):
- lead optimization, desirability functions, 163–164
- multiparameter lead optimization, property profiles, 168–169
Quantitative structure-activity relationship (QSAR):
- activity cliff quantification, 210–211
- chemical arrays, 181–182
- lead optimization, multiobjective processes, 156–157
- metabolic lability model, 253–256
- DGAT1 inhibitors, 258–262
- omics applications, 86
- partial least square method, 85–86
- aminergic GPCR inhibitory activity data, 86–87
- atomic colorings:
  - molecular modeling, 91–93
  - regression coefficient matrix, 90–91
- LPLS analysis, 91
- LPLS ligand and protein descriptors, 87–88
- L-shaped PLS architecture, 89–90
- safety modeling:
  - cardiac liability risk reduction, 271–272
  - machine learning, 270–271
  - predictive secondary pharmacology vs., 274–276
- in vitro endpoints, 269–270
- similarity principle, 207
Quantitative structure-property relationship (QSPR), metabolic lability model, 257–262
DGAT1 inhibitors, 258–262
Query compounds:
- safety QSAR modeling, 270–271
- visualization in SAR, 221–222
Query engine, Mobius CIDB system, 333
Random forest model, building and testing, 9–11
Random walk models, lead optimization, 191–194
Rapid Overlay of Chemical Structures (ROCS) program:
  antitarget activity analysis, 230
  molecular similarity models, 39–45
  structure activity relationships, three-dimensional ligand-based relationships, 225–227
RDKit chemoinformatics programming library, predictive modeling and, 2–31
Reactive Metabolite Warning System (RMWS), toxicity studies, 277–278
Read-across analysis, toxicity studies, 274
RECAP methodology:
  maximum common substructure, 208
  seed compounds, chemical arrays, 199–201
Receiver operating characteristic (ROC), 377–378
Recombinant protein assays, multiobjective lead optimization, 156–157
Reduced graphs, molecular scaffolds, 208
Reference (probe) molecule:
  asymmetric similarity, 366–370
  molecular similarity measurements, 375–376
Regression coefficient matrix, L-shaped partial least squares analysis, atomic colorings, 90–91
Regression models, L endo-and exo-LPLS, chemogenomics analysis, 89–90
Rescaffolding, structure-activity relationships transfer, 223–227
  three-dimensional ligand-based approaches, 225–227
  three-dimensional protein-based approaches, 227
Retrospective analyses, lead optimization, 169, 171–173
Reverse virtual screening, 135–137
RGate 2003+ CIDB, 327
R-groups:
  lead optimization:
    activity cliffs exploration, 214–216
    matched molecular pairs (MMPs), 212–214
    seed compound detection, 199
    self-avoiding random walks, 191–192
    visualization techniques, 216–222
Ring assemblies:
  BICLAIM-space searches, 311–314
  chemotype detection, chemical arrays, 196–198
ROCK system, chemical arrays, 202
Rollout systems, chemoinformatics databases, apostle approach, 338
Root-mean-square deviation (RMSD):
  predictive model performance evaluation, 8
  random forest model construction, 10–11
Rotatable bonds (RotB), multiparameter lead optimization, 158–159
  desirability functions, 163–164
Rotated factorial design, lead optimization, 151–152
Rough set theory (RST):
  chemically induced gene expression data, 57–61
  classification accuracy and quality, 56–57
  drug-induced gene expression, phospholipidosis, 65–77
  dataset characteristics, 68–71
  D-reduct determination, 68, 71–72
  preliminary rule generation, 72–75
  rule simplification-attribute value reduction, 75–77
  rule generation, 63–64
R statistics program:
  predictive modeling, 2–31
  molecular similarity applications, 12–15
  random forest model construction, 9–11
Rule-based methods, 51–52
Safety profiles:
  biological data, 279–283
  chemical structure and available data, 272–278
  combination of evidence, 278–279
  data-driven modeling, 269–283
  future research issues, 285–286
  preclinical data, 282–283
  predictive secondary pharmacology, 274–276
  quantitative structure activity relationships:
    cardiac liability risk reduction, 271–272
    machine learning, 270–271
    in vitro endpoints, 269–270
  read-across analysis, 274
  warning systems in, 276–278
SaliExplorer program, activity cliff analysis, 215–216
SAR Map technique, visualization in SAR analysis, 216–222
Scaffold compounds. See also Rescaffolding
  chemotype detection, 196–198
  structure-activity relationship, 207–208
Scaffold hopping potential, virtual screening performance, 118–119
Scaffold trees, molecular scaffolds, 208
Scores matrix, three-dimensional ligand-based relationships, 226–227
Scoring profiles, multiparameter lead optimization, probabilistic scores, 164–165
SCRUM, safety modeling, 283–285
Secins, cytohesin inhibitors, 124–125
Seed-array scatter plots, chemical arrays, 182–186
Seed compound detection, chemical arrays, 196, 198–201
Selectivity cliffs, 216
self-avoiding random walk (SAW), lead optimization, 191–194
Self-organizing maps (SOMs), visualization in SAR and, 219–222
Semi-naïve Bayesian models (SNBMs), 144–145
Sequel language, 323
Sequential screening, virtual screening, 121
Serendipity, drug design, 35–36
Side effect pairing, 34
Sildenafil, protein-ligand binding, 38–39
Similarity ensemble approach (SEA):
naïve Bayesian models, class comparisons, 143
safety modeling, predictive secondary pharmacology, 275–276
Similarity functions, molecular similarity:
chemical graph-based similarity functions, 361–363
set-based similarity functions, 357–361
vector-based and function-based functions, 363–365
Similarity fusion, molecular similarity analysis, 372
Similarity matrices, molecular similarity techniques, 376–377
Similarity principle. See also Molecular similarity analysis
Simplex optimization, lead optimization, 153–154
Simplified molecular input line entry system (SMILES):
chemotype detection, chemical arrays, 195–198
matched molecular pairs (MMPs), 212–214
molecular similarity applications, 98–100
two-dimensional analysis, 101
Singular value decomposition, chemogenomics analysis, 89–90
Site of metabolism (SOM) predictions:
CYP enzymes, 247
metabolic lability model, DGAT1 inhibitors, 257–262
Small molecule drug development, molecular similarity and, 39–45
SMARTS notation, matched molecular pairs (MMPs), 212–214
Software technology:
chemoinformatics databases, quality controls, 336
Mobius CIDB system, 335
off-the-shelf software, 338
safety modeling, 283–285
Solvated states, protein-ligand binding and, 37
Source code listings, predictive models, 20–30
Speed hazards, chemoinformatics databases, 337
Spiral view techniques, visualization in SAR, 221–222
Splitting operations, BICLAIM-space searches, 312–314
Spotfire Decision Site, 328
Stability analysis, molecular fingerprinting:
future research issues, 110–111
isotere generation, WABE program, 101–102
results, 103–110
Tanimoto similarity, 102
two-dimensional methods, 100–101
Standard deviations, Tanimoto similarity, 102
Statistical independence, molecular similarity, 374–375
Stepwise analysis, metabolic lability model, 251–256
Stochastic neighborhood embedding (SNE), visualization in SAR and, 219–222
Structural alerts, toxicity analysis, 273–274
Structural complexity (SC), chemical arrays, 195–198
Structure-activity landscape analysis, 326
Structure-activity landscape index (SALI):
activity cliff quantification, 210–211, 215–216
molecular similarity, 383–384
antitarget activity analysis, 229–230
Structure-activity relationship index (SARI),
activity cliff quantification, molecular similarity, 384
Structure-activity relationships (SARs). See also Quantitative structure-activity relationship (QSAR)
ADMET models, 246–247
Boehringer Ingelheim chemoinformatics case study, 305–309
chemical arrays, 180–181, 201–203
drug design, 35–47
drug-induced gene expression, phospholipidosis, 66
lead optimization:
activity cliffs:
exploration, 214–216
outlier investigation, 209–210
quantification, 210–211
antitarget activity, 227–232
activity hotspot identification and application, 228–230
hERG and CYP3A4 inhibition, 230–231
project integration, 232
transfer mechanisms, 227–228
future research issues, 232–233
matched molecular pairs, 211–214
Structure-activity relationships (SARs). (Continued)
molecular scaffolds, 207–208
overview, 205–206
pharmaceutical industry applications, 222–223
privileged substructures, 208–209
resciffolding, 223–227
three-dimensional-ligand-based approaches, 225–227
two-dimensional approaches, 224–225
resciffolding, 223–227
molecular similarity principle, 207
visualization support, 216–222
molecular similarity and, 97–100
multiobjective lead optimization, 154–157
future research issues, 176–177
 naïve Bayesian models, 132–134
molecular similarity combined with, 138–139
virtual screening, 121–123
Structure-activity similarity (SAS) maps:
molecular similarity applications, 384–387
visualization in SAR and, 221–222
Structure-based ADMET models, development and applications, 246–247
Structure-based virtual screening (SBVS):
correlation methods, molecular similarity and, 375–376
definition and classification, 113–114
ecto-5 ’-nucleotidase first-in-class inhibitor, 125–128
performance evaluations, 117–119
practical applications, 116–117
Structured Query Language (SQL), 323
Mobius CIDB system, 333
Structure-property relationships (SPRs), Bohering Ingelheim chemoinformatics case study, 308–309
Structure similarity maps, visualization in SAR analysis, 219–222
Substructures:
BICLAIM-space searches, 311–314
privileged substructures, 208–209
structural alerts, toxicity studies, 273–274
Succession planning, chemoinformatics database systems, 339
Superfluous attributes:
drug-induced gene expression, 75–77
rough set theory, 61–63
Support systems, chemoinformatics database systems, 339
Support vector machine (SVM) modeling, 124–125
Surface interactions, molecular similarity models, 39–45
Surflex-Dock molecular docking, protein-ligand-binding and, 38–39
Surflex-QMOD approach, three-dimensional QSAR, activity prediction, 45–47
Surflex-Sim approach, molecular similarity models, 39–45
Tadalafil, protein-ligandbinding, 38–39
Tanimoto similarity:
fingerprint method comparisons with, 104–110
Huuskonen training set, 13–15
molecular fingerprinting techniques:
expected values, 110–111
two-dimensional applications, 102
set-based similarity functions, 360–361
Target fishing:
 asymmetric similarity, 350
 naïve Bayesian models:
 phenotype comparison, chemical and biological activity space, 141–143
 reverse virtual screening, 135–137
Temporal analysis:
 lead optimization, chemical arrays, 186–191
metabolic lability model, 255–256
Temporal partitioning, molecular similarity model, 45–47
Test-Driven Development (TDD), safety modeling, 284
Testicular Toxicity Warning System (TTWS), 278
Three-dimensional activity landscape models, activity cliff analysis, 215–216
Three-dimensional fingerprints, molecular similarity, set-based representation, 352–353
Three-dimensional ligand-based models:
ADMET models, 246–247
structure-activity relationships, 225–227
Three-dimensional protein-based techniques, structure-activity relationships, 227
Three-dimensional QSAR:
ADMET models, 246–247
fingerprint techniques, 98–100
physically realistic activity prediction, 45–47
Three-dimensional similarity, computational techniques, 370–371
Three-point pharmacophore fingerprints, structure-activity relationships, 225–227
Top-down processing, chemoinformatics databases, 338
Topliss Tree, lead optimization, 150–154
Topoisomer searching, rescaffolding, 224–225
Topological frameworks:
  molecular scaffolds, 208
  rescaffolding, 224–225
Toxicity studies:
  biological data, 279–283
  chemical structure and available data, 272–278
  customer interface, 283–285
  data-driven modeling, 269–283
  preclinical data, 282–283
  predictive secondary pharmacology, 274–276
  read-across analysis, 274
  structural alerts in, 273–274
  in vitro analysis, chemical structure links, 269–272
  warning systems in, 276–278
Traffic light visualization, multiparameter lead optimization, 160–161
  probabilistic scores, 164–165
Training programs, chemoinformatics database systems, 339
Training set molecules, predictive model performance vs., 12–15
Tree-based fingerprints, two-dimensional analysis, 101
t–u plots:
  activity cliff quantification, 210–211
  visualization in SAR, 217–222
Tversky similarity, asymmetric similarity, 367–370
Two-dimensional representation:
  chemical graph-based representations, 353–354
  computational techniques, 370–371
  set-based representation, 352–353
  naïve Bayesian models, comparisons with, 138–139
  rescaffolding, non-fingerprint techniques, 224–225
Unbound states, protein-ligand binding, 37
Union sets, set-based similarity functions, molecular similarity, 357–359
Unity fingerprints, 254–256
Upregulated scores, drug-induced gene expression, phospholipidosis, 68–71
Urotensin II receptor (UTR), virtual screening, 122–123
Usage facilitation, safety profiling and, 284–285
User interface, chemoinformatics databases, 336
Validation, molecular similarity techniques, 377–378
Van der Waals interactions, protein-ligand binding, 37–39
Variable-length fingerprints, molecular similarity
  set-based representation, 352–353
Vector-based representation, molecular similarity, 354–356
  similarity functions, 363–365
Virtual combinatorial libraries, 297–299
Virtual screening (VS):
  applications, 116–117, 123–128
  first-in-class inhibitor, ecto-5’-nucleotidase, 125–128
  multifunctional protein inhibitors, 124–125
  definition and classification, 113–114
  high-throughput screening vs., 119–121
  molecular drug targets, 121–123
  naïve Bayesian models, 134–137
  performance evaluation, 117–119
Visualization techniques, structure-activity relationships, 216–222
Volumetric functions, molecular similarity models, 41–45
WABE program, isostere generation, molecular fingerprinting applications, 101–102
Warning systems, toxicology analysis and, 276–278
Water partition coefficient, multiparameter lead optimization, 158–159
Weighted desirability scoring, temporal analysis, lead optimization, 186–191
Weighted representations, molecular similarity, 356–357
Weight of evidence (WoE), toxicity analysis, 279–280
Workflow systems, 294–295, 315–317
z-scales, L-shaped partial least squares analysis:
  heat mapping, regression coefficient matrix, 92–93
  ligand and protein descriptors, 87–88
z-score:
  fingerprint method comparisons, 106–110
  Tanimoto similarity, 102