3
Data compilation and Curation

Towards the objective of obtaining drug features that could be used for integrating in a model for prediction of side effects, we compiled and curated empirical data from existing databases. Specifically, we used DrugBank 3.0 [Knox et al., 2011] and SIDER [Kuhn, Campillos, Letunic, Jensen, & Bork, 2010] that provide complementary information of drugs, their targets and phenotypic side effects.

3.1 INTEGRATION OF DRUGBANK 3.0 AND SIDER2

For the purpose of addressing questions aligned with objectives in this thesis, we integrated data of drugs, proteins they are known to target and their side effects**.**

DrugBank: DrugBank provides one of the most comprehensive information on drugs and their targets. This data of drug-target relationships include: DrugBank identification number (DBID) of drugs (an alphanumeric code), target's name and its gene sequence. This resource provides the target-related data classified across classes such as 'Drug Target Sequences', 'Drug Enzyme Sequences', 'Drug Carrier Sequences' and 'Drug Transporter Sequences'. Each of these data are segregated into Approved, Small Molecule, Biotech, Experimental, Neutraceutical, Illicit, Withdrawn and Investigational. We used the 'Drug Target Sequences' information that encodes drug-target relationship in terms of UniProt ID and the name of protein, DNA sequence and DrugBank IDs of drugs linked to the target. These details were extracted into separate files from the source data (DrugBank 3.0) using a python script. The final data thus compiled using source file (obtained in March 2014) comprised 5788 drugs and their 3520 targets.

SIDER: The information on drug side effects was obtained from SIDER2 (obtained in March 2014) [Kuhn, Campillos, Letunic, Jensen, and Bork, 2010]. This data is presented with the following structure: STITCH ID, UMLS concept ID, Drug name, Side effect name, MedDRA concept type (LLT- lowest level term, PT- preferred term), UMLS concept id corresponding to MedDRA term, MedDRA side effect name. For our analysis, we extracted information of drug name, UML concept ID and MedDRA side effect names. A total of 101852 drug side-effect relationships were obtained comprising of 996 unique drugs and 4192 distinct side-effects. Clearly, the information of drug side effects is limited as compared to drug-target interaction information available from DrugBank.

Integration: To address some of the questions asked as part of the thesis objectives (Chapter 5 and Chapter 6) we needed to have composite data of drugs, their targets and side effects. Since these databases did not use unique IDs for drugs, we mapped drugs across these datasets to find the data intersection. We implemented following strategy for integration of data from DrugBank 3.0 and SIDER2 to obtain the desired composite dataset (Figure 3.1). This strategy is also reported as part of data compilation of [Kanji, Sharma, and Bagler, 2015] and [Sharma, 2015].

In Stage-I, we identified drugs that were trivially common by mapping generic names. In Stage-II, using a multi-step process we resolved the aliases of a drug to match more drugs. Thus we arrived at an integrated dataset of drugs with target and side effects characterization.

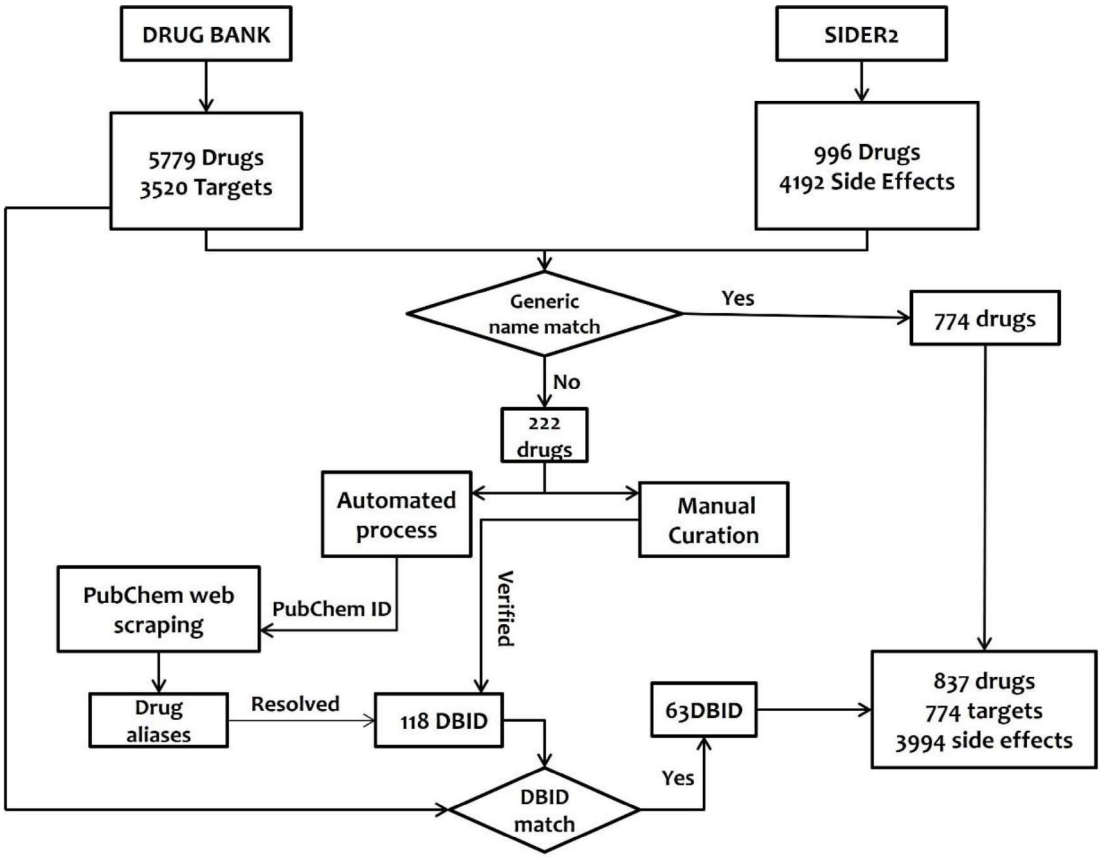

Figure 3.1 : The procedure implemented for integration of drugs, their targets and side effects. Figure adapted from [Sharma, 2015].

Stage-1: The generic names of 774 drugs out of the total 996 drugs listed in SIDER could be matched exactly with that in DrugBank. The specific details of side effects were then linked with data obtained from SIDER (meddra_adverse_effects.tsv).

Stage-II: For the remaining 222 drugs, a two-step procedure was used for finding out aliases of drugs from SIDER, followed by their detailed verification. The first step was performed using an automated process for obtaining drug alias information by scraping PubChem.

1. The Pubchem IDs of 222 unresolved SIDER drugs was used as an input data.
2. A python script was used for scraping the Pubchem website.
3. The drug name, along with its aliases and chemical formula were extracted.
4. The drug name thus obtained were searched in DrugBank. The drug was putatively linked to a DrugBank ID in case of direct match.
5. The chemical formula in PubChem and DrugBank were compared.
6. For a successful match, the DrugBank ID was matched with the query drug.

Following were the statistics at the end of this process:

- Number of drugs in SIDER2: 996
- Number of drugs for which trivial match with generic name could not be obtained: 222
- Number of drugs resolved using DBID (out of 222): 118

Following protocol was implemented for verification of the mapping obtained with the automated protocol:

- Starting with the Pubchem ID and DBID of a drug (for each of the 118 drugs identified), details available from PubChem as well as DrugBank were compared manually.
- The chemical structures were used as the basis for comparison.
- The accuracy of the automated procedure was cross-checked by manually going through the drugs with mismatch.

The flow diagram in Figure 3.2 depicts this process starting with the data from DrugBank and SIDER2 leading to a composite dataset of 837 drugs, 774 targets and 3994 side effects.

3.2 STATISTICS OF DRUGBANK 3.0 AND SIDER

The phenotypic effects arising from action of a drug could be seen as a response due to drug's interactions with cell's molecular machinery. The data from DrugBank reflects such interactions reported from experimental assays. They comprise of four type of FDA approved drugs: small molecules, FDA approved biotech (protein or peptide), nutraceutical (Vitamin, metabolites), and experimental (de-listed, illicit, and enzyme inhibitor). These data have been compiled from various sources such as textbooks, journals, databases, text mining tools and web-based programs. Further, these were manually cross verified by experts like physicians, bioinformaticians and biochemists [Wishart et al., 2006].

The matrix of drug-target associations was found to be sparse with only 2901 interactions (which is less than 1% of the maximum interactions possible interactions 647,838). This data could be seen from two perspectives: (a) by looking for drugs and the number of targets they bind to, and (b) by looking for targets and the number of drugs they are bound by. Figure 3.2, which depicts these details in the form of cumulative distributions, suggests that while every target is bound by at least one drug, no target is bound by more than 57 drugs (Figure 3.2(a)). Similarly, every drug binds to at least one target, and the largest number of targets that a given drug binds to is 34 (Figure 3.2(b)). Thus, both data show heterogeneous nature with 'on an average 3.46 drugs binding to a target' and 'on an average 3.74 number targets bound by a drug'.

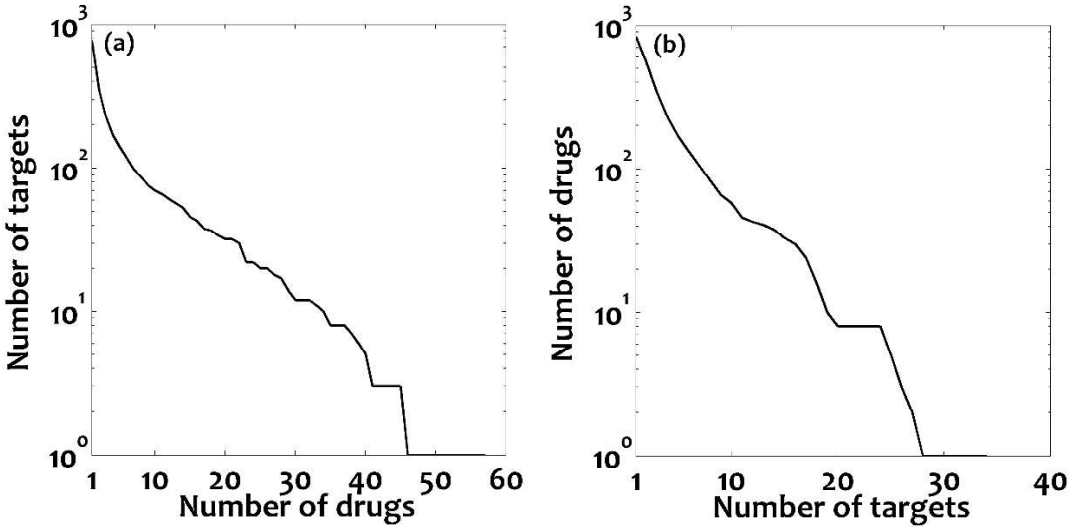

Figure 3.2 : (a) Number of targets that are bound by ≥d drugs. (b) Number of drugs that bind ≥t targets.

Similarly the data from SIDER2 as well as SIDER4 [Kuhn, Letunic, Jensen, and Bork, 2016] was analyzed from two perspectives: (a) number of side effects caused by drugs, and (b) number of drugs that are causing side effects. The sparseness of drug-side effects matrix obtained from SIDER2 was 2.64%, with 88109 relationships out of 334,2978 maximally possible. The newer

version SIDER4 had enhanced data of 1430 drugs and 5868 side effects with around 43% and 40% increase in data, respectively. As seen in Figure 3.3(a), every drug in SIDER4 was presented with at least one side effect and the drug with largest number of side effects had 1207 side effects. The number of drugs having more than a certain number of side effects falls, with no drug presented with more than 769 side effects as depicted in Figure 3.3(b).

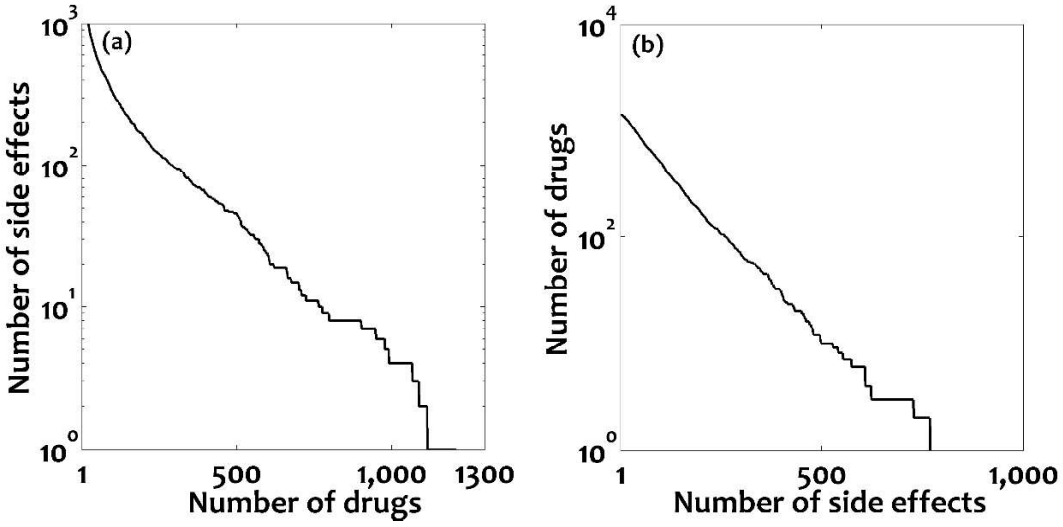

Figure 3.3 : (a) Number of side effects that are caused by per ≥d drugs. (b) Number of drugs are associated with ≥s side effects.

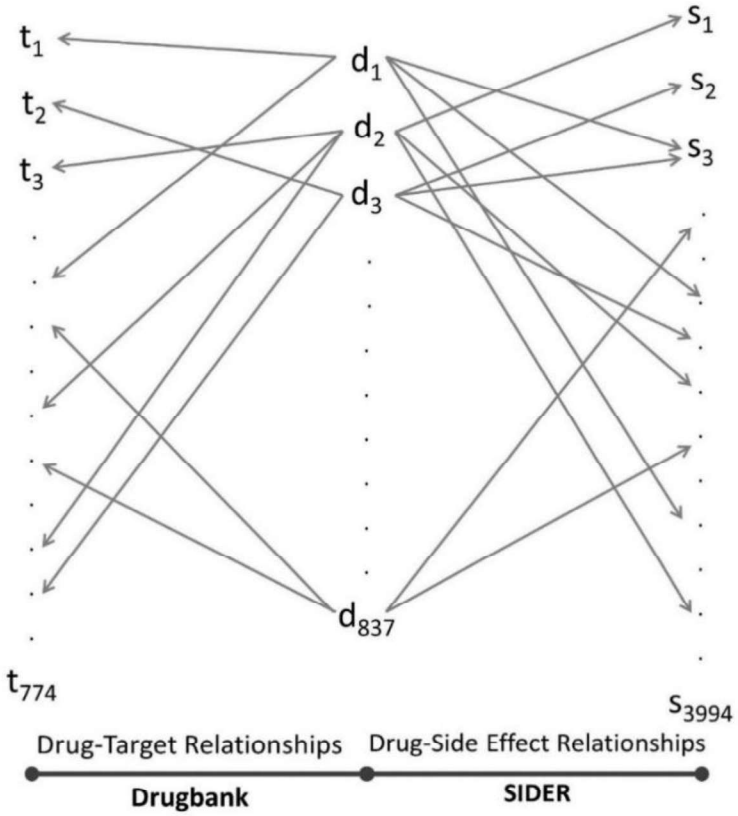

Figure 3.4 : Illustration of the Drug-Target-Side tri-partite relationships. The drug-target regulatory associations (DrugBank) and drug-side effect data (SIDER) was merged to create a composite dataset. This integrated dataset was used for investigations presented in Chapter 5 and Chapter 6.

Figure 3.4 represents relationships between key entities used in this thesis (drugs, targets and side effects). Drug is depicted as a central entity which is reported to 'bind/interacts' with certain targets. A drug could bind to one more targets. On the phenotypic side, the drug could be reported to 'cause' certain side effects. A given drug could be causally associated with one or more side effects. Such integration of data from DrugBank and SIDER facilitates computational investigations to answer questions that are central to objectives of this thesis.

3.3 CHEMICAL DESCRIPTORS

Drug molecules are represented with various formats which can be to compute chemical properties as well. In this work, Discovery Studio 4.0 was used to compute drugs properties. In subsequent sections, a brief description of chemical representations and properties of drugs have been discussed.

3.2.1 Computational representation of molecules

Computational representation of molecules are encoded with various formats (such as SMILES, InChI, IUPAC etc) as well as using their 2D and 3D structural representation. For example, drug Levobunolol is represented in various formats as shown in Table 3.1 and Figure 3.2.

Table 3.1 SMILES, InChI and IUPAC representations of Levobunolol.

SMILES	InChI	IUPAC
CC(C)(C)NC[C@H](O)COC1=CC=CC2=C1CCCC2=O	InChI=1S/C17H25NO3/c1-17(2,3)18-10-12(19)11-21-16-9-5-6-13-14(16)7- 4-8-15(13)20/h5-6,9,12,18-19H,4,7-8,10-11H2,1-3H3/t12-/m0/s1	(S)-5-{[3-(tert-Butylamino)-2-hydroxypropyl]oxy}-3,4-dihydronaphthalen-1(2H)-on

In this thesis, we have used SMILES, which encode any molecule as a string, as well as SDF representations. In 3D structure representation, atomic coordinates (X, Y and Z) are provided.

3.2.2 Chemical properties

Chemical properties of drugs were computed using 2D and 3D structures using Discovery Studio 4.0 [Todeschini and Consonni, 2000; van de Waterbeemd and Testa, 2003]. It classifies 2D and 3D chemical properties under various broad. A total of 271 and 56 features representing 2D and 3D properties, respectively, were used for our studies. Following is a summary of these descriptors.

Classes of 2D properties

Following are the broad classes for 2D properties that are measured with 2D conformation of molecules. Complete list of properties for each of the following classes is provided in Tables 3.2.

Electrostatic property: These describes electrostatic details based on charge distributions in the molecules and comprises of 156 descriptors.

Molecular property: These encompass information of 102 descriptors with details of type of inter-atoms bonds, number of rings within, Lipinski's features etc.
Molecular surface area: These describe availability of a molecule when it is introduced with solvent and comprises of 13 descriptors.

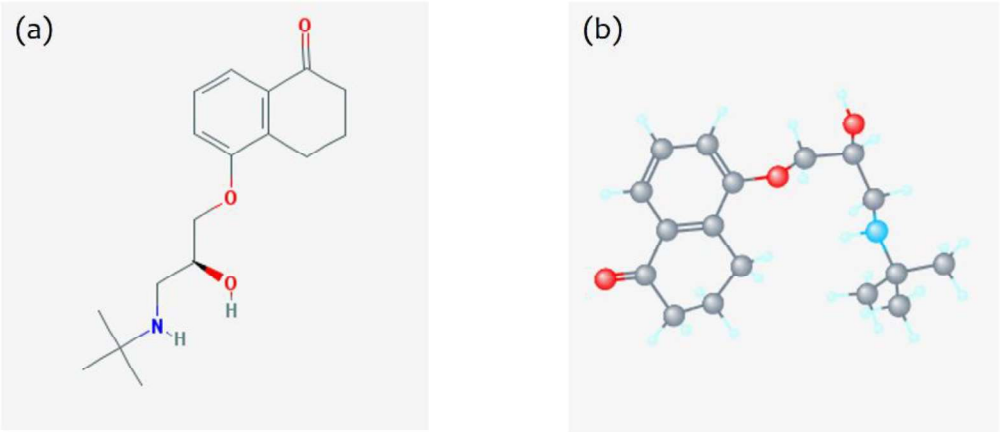

Figure 3.5 (a) 2D and (b) 3D structure representation of Levobunolol. Images were obtained from PubChem.

Table 3.2 List of 2D chemical classes and their descriptors

Electrostatic property	ES Count aaaC (1), ES Count aaaCH (2), ES Count aaN (3), ES Countaa NH (4), ES Count aao (5), ES Count aaS (6), ES Count aasC (7), ES Count aaSe (8), ES Count aasN (9), ES Count dCH (10), ES Count ddC (11), ES Count ddsN (12), ES Count ddssS (13), ES Count ddssSe (14), ES Count dNH (15), ES Count dO (16), ES Count dS (17), ES Count dsCH (18), ES Count dSe (19), ES Count dN (20), ES Count dssC (21), ES Count dssS (22), ES Count dssSe (23), ES Count dsssP (24), ES Counts AsH2 (25), ES Count sBr (26), ES Count sCH (27), ES Count sCl (28), ES Count sF (29), ES Count sGeH (30), ES Count sI (31), ES Count sLi (32), ES Count sNH2 (33), ES Count sNH3 (34), ES Count sOH (35), ES Count sPbH3 (36), ES Count sPH2 (37), ES Count ssAsH (38), ES Count ssBe (39), ES Count ssBH (40), ES Count ssCH2 (41), ES Count sSeH (42), ES Count ssGeH2 (43), ES Count sSH (44), ES Counts SiH3 (45), ES Count ssNH (46), ESCount ssNH2 (47), ES Count sSnH (48), ES Count ssO (49), ES Count ssPbH2 (50), ES Count ssPH (51), ES Count ssS (52), ES Count sssAs (53), ES Count sssB (54), ES Count sssC (55), ES Count sssdAs (56), ES Count ssSe (57), ES Count sssGeH (58), ES Count ssSiH2 (59), ES Count sssN (60), ES Count sssNH (61), ES Count ssSnH2 (62), ES Count sssP (63), ES Count sssPbH (64), ES Count ssssB (65), ES Count ssssBe (66), ES Count ssssC (67), ES Count ssssGe (68), ES Count sssSiH (69), ES Count sssN (70), ES Count sssSnH (71), ES Count ssssPb (72), ES Count sssssAs (73), ES Count ssssSi (74), ES Count ssssSn (75), ES Count sssssP (76), ES Count tCH (77), ES Count tN (78), ES Count tsC (79), ES Sum aaaC (80), ES Sum aaCH (81), ES SumaaN (82), ES Sum aaNH (83), ES Sum aaO (84), ES Sum aaS (85), ES Sum aasC (86), ES Sum aaSe (87), ES Sum aasN (88), ES Sum dCH2 (89), ES Sum ddC (90), ES Sum ddsN (91), ES Sum ddssS (92), ES Sum ddssS (93), ES Sum dNH (94), ES Sum dO (95), ES Sum dS (96), ES Sum dsCH (97), ES Sum dSe (98), ES Sum dsN(100), ES Sum dssC (101), ES Sum dssS (101), ES Sum dssSe (102), ES Sum dsssP (103), ES Sum sAsH2 (104), ES SumsBr (105),

	ES Sum sCH3 (106), ES Sum sCl (107), ES Sum sF (108), ES Sum sGeH3 (109), ES SumsI (110), ES Sum sLi (111), ES Sum sNH2 (112), ES Sum sNH3 (113), ES Sum sOH (114), ES Sum sPbH3 (115), ES Sum sPH2 (116), ES Sum ssAsH (117), ES Sum ssBe (118), ES Sum ssBH (119), ES Sum ssCH2 (120), ES Sums she (121), ES Sum ssGeH2 (122), ES Sum sSH (123), ES Sum sSiH3 (124), ES Sum ssNH (125), ES Sum ssNH2 (126), ES Sum sSnH3 (127), ES Sum ssO (128), ES Sum ssPbH2 (129), ES Sum ssPH (130), ES Sum ssS (131), ES Sum sssAs (132), ES Sum sssB (133), ES Sum sssCH (134), ES Sum sssdAs (135), ES Sum ssSe (136), ES Sum sssGeH (137), ES Sum ssSiH2 (138), ES Sum sssN (139), ES Sum sssNH (140), ES Sum ssSnH2 (141), ES Sum sss (142), ES Sum sssPbH (143), ES Sum ssssB (144), ES Sum ssssBe (145), ES Sum ssssC (146), ES Sum ssssGe (147), ES Sum sssSiH (148), ES Sum ssssN (149), ES Sum sssSnH (150), ES Sum ssssPb (151), ES Sum sssssAs (152), ES Sum ssssSi (153), ES Sum ssssSn (154), ES Sum sssssP (155), ES Sum tCH (156)
Molecular property	Aplo (157), CoordDimension (158), Formal Charge (159), Gasteiger Charges (160), IsChiral (161), Molecular Solubility (162), pKa Atom (163), QED ALOGP (164), QED AROM (165), QED MW (166), QED PSA (167), QED ROTB (168), QED Unweighted (169), SAscore (170), SAscore Complexity (171), SAscore Fragments (172), VSA AlogP (173), VSA TotalArea (174), VSA Atomic Areas (175), VSAMR (176), VSA Partial Charge (177), HBA Count (178), NPlusO Count (179), Num AromaticBonds (180), Num AromaticRings (181), Num AtomClasses (182), Num Atoms (183), Num AtropisomerCenters (184), Num AxialStereoCenters (185), Num Bonds (186), Num BridgeBonds (187), Num BridgeHeadAtoms (188), Num ChainAssemblies (189), Num Chains (190), Num ComplexedFragments (191), Num CustomData (192), Num DativeBonds(193), Num DoubleBonds (194), Num EnhancedStereoAtoms (195), Num ExplicitAtoms (196), Num ExplicitBonds (197), Num ExplicitHydrogens (198), Num Fragments (199), Num HAcceptors (200), Num HAcceptorsLipinski (201), Num HDonors (202), Num HDonorsLipinski (203), Num HomologySGroups (204), Num HydrogenBonds (205), Num Hydrogens (206), Num Isotopes (207), Num LinkAtoms (208), Num MacroChains (209), Num MacroResidues (210), Num MarkushBonds (211), Num MesoStereoAtoms (212), NumMetalAtoms (213), Num MixtureComponents (214), Num Mixtures (215), Num Monomers (216), Num MulAtoms (217), Num NegativeAtoms (218), Num PiBonds (219), Num Polymers (220), Num PositiveAtoms (221), Num PseudoStereoAtoms (222), Num QueryAtoms (223), Num QueryBonds (224), Num RepeatUnits (225), Num RGroupFragments (226), Num RingAssemblies (227), Num RingBonds (228), Num RingFusionBonds (229), Num Rings (230), NumRings3 (231), Num Rings4 (232), Num Rings5 (233), Num Rings6 (234), Num Rings7 (235), Num Rings8 (236), Num Rings 9Plus (237), Num RotatableBonds (238), Num SCSRLocalTemplates (239), Num SCSRSequenceAtoms (240), Num SCSRTemplates (241), Num SGroups (242), Num SingleBonds (243), Num SpiroAtoms (244), Num StereoAtoms (245), Num StereoBonds (246), Num Superatoms (247), Num TerminalRotomers (248), Num TripleBonds(249), Num TrueAlleneStereoCenters (250), Num TrueAtropisomerCenters (251), Num TrueStereoAtoms (252), Num UnknownPseudoStereoAtoms (253), Num UnknownStereoAtoms (254),

	Num UnknownStereoBonds (255), Num UnknownTrueStereoAtoms (256), Num V3000Templates (257), Organic Count (258)
Molecular surface area	Molecular FractionalPolarSASA (259), Molecular FractionalPolarSurfaceArea (260), Molecular PolarSASA (261), Molecular PolarSurfaceArea (262), Molecular SASA (263), Molecular SAVol (264), Molecular SurfaceArea (265), BIC (266), CHI 0 (267), CHI 1 (268), CHI 2 (269), CHI 3 C (270), CHI 3 CH (271)

Classes of 3D properties

Similar to 2D properties, 3D properties of molecules are presented in following broad classes that are measured with their 2D conformation. Complete list of properties for each of the following classes is provided in Tables 3.3.

Dipole moment: This class comprises of 4 descriptors with quantification of the polarity of a molecule

Jurs descriptor: These 34 descriptors enumerate properties of the molecule based on the electronic charge present with categories of partial positive and negative, including at the level of atoms.

Energy: These three descriptors encode the total, minimized and strain energy of a molecule.

Principal moment: These five descriptors enumerate magnetic moment, its components along three axes and the radius of gyration.

Molecular surface area: These four descriptors specifies the solvent availability of a molecule under different categories.

Shadow length: Projection length of the 3D molecule along different axis and planes are provided with 10 descriptors.

Table 3.3 List of 2D chemical classes and their descriptors

Dipole moment	Dipole mag (1), Dipole X (2), Dipole Y (3), Dipole Z (4)
Jurs descriptors	Jurs DPSA 1 (5), Jurs DPSA 2 (6), Jurs DPSA 3 (7), Jurs FNSA 1 (8), Jurs FNSA 2 (9), Jurs FNSA 3 (10), Jurs FPSA 1 (11), Jurs_FPSA_2 (12), Jurs FPSA 3 (13), Jurs PNSA 1 (14), Jurs PNSA 2 (15), Jurs PNSA 3 (16), Jurs PPSA 1 (17), Jurs PPSA 2(18), Jurs PPSA 3(19), Jurs RASA (20), Jurs RNCG (21), Jurs RNCS (22), Jurs RPCG (23), Jurs RPCS (24), Jurs RPSA (25), Jurs SASA (26), Jurs TASA (27), Jurs TPSA (28), Jurs WNSA 1 (29), Jurs WNSA 2 (30), Jurs WNSA 3 (31), Jurs WPSA 1 (32), Jurs WPSA 2 (33), Jurs WPSA 3 (34)
Energy	Energy (35), Minimized Energy (36), Strain Energy (37), Clean Energy (38)
Principal moment	PMI mag (39), PMI X (40), PMI Y (41), PMI Z (42), Rad Of Gyration (43)
Molecular surface area	Molecular 3D PolarSASA (44), Molecular 3D SASA (45), Molecular 3D SAVol (45), Molecular Volume (46)

Shadow length	Shadow nu (47), Shadow Xlength (48), Shadow_XY (49), Shadow_XYfrac (50), Shadow XZ (51), Shadow XZfrac (52), Shadow Ylength (53), Shadow YZ (54), Shadow YZfrac (55), Shadow Zlength (56)

All computations done as part of studies reported in this thesis were performed on Dell Precision T5610 workstations (*Charaka, Sushruta*) of the Complex Systems Laboratory, IIT Jodhpur.

…