



CoSSDb: A Database of Co-crystallized Ligand Sub-structures for Anticancer Lead Designing & Optimization

Om Prakash and Feroz Khan^(✉)

Computational Biology Unit, CSIR-CIMAP, Lucknow 226015, India
f.khan@cimap.res.in

Abstract. The Discovery of the novel optimized structures of small molecules for selective targeting is one of the challenging tasks in drug designing. Bioisosteres are the key components of the lead compound, which provide hidden power to the compound scaffold for selective targeting. We are presenting a database, named CoSSDb which stands for Co-crystallized Sub-Structure Database. The CoSSDb contains ligand sub-structures as possible bioisosteres. extracted from PDB files, available in Protein Data Bank. Sub-structures were extracted through an algorithm, which utilizes the location of atoms in the 3D domain of the complex ligand & protein. It processes the relative positioning of atoms for demarcation of the influential part of the ligand, which interacts with macromolecule and provides potency to that ligand for binding with a specific binding pocket of the protein. The algorithm was used to extract sub-structures from the ligands co-crystallized with proteins involved in cancer. About 7721 x-ray crystallography PDB files were processed, and 654 non-redundant substructures were identified. These sub-structures will be useful during designing & optimization of novel ligands for selective targets. The database is freely accessible at '<https://opticket49.wixsite.com/substructdb>'.

Keywords: Cancer · Bioisostere · Ligand · PDB · Sub-Structure

1 Introduction

The selective targeting of proteins by ligand in cellular processes is one of the primary challenges in toxicology and new drug discovery. As a remedy to this challenge, the structure of ligands or compounds that are known to interact with specific cell signalling proteins is being explored. Understanding of structural components are must for performing studies on selective targeting, drug design, lead editing, and lead optimization etc. Some of the previously known methods for lead optimization for anticancer activity are as: Experiment based hit & trials, sub-structure & structure alerts-based lead identification, and fragment-based lead optimization etc. With existing databases, significance of our database can be seen in terms of availability of 'authentic side chains' for lead optimization, and selectivity towards the specific set of residues [1]. Because the role of sub-structural components is well established, several attempts to identify

structural components for this purpose have been made in the past. The significance of sub-structure has been explained in the drug-target network [2]. The importance of structural underpinnings for selective targeting has also been proven [1]. There have been numerous ways explained for discovering sub-structures that act as bioisosteres. These earlier methods used structure comparison processes based on structural similarity, fingerprinting, and QSAR modelling, among others, to identify sub-structure or structural warnings [3]. To find substructures in previous studies, structural comparison was employed. European REACH utilises a weight-of-evidence approach to identify bio-accumulative substances. One of the components in REACH's weight of evidence is structural alarm detection based on quantitative structure activity relationship (QSAR). It's worth noting that in this work, QSARs were used to extract structural alerts from compounds [3]. For structural warnings, statistical QSAR models based on structural traits were created to indicate likely chemical dangers [4]. To reduce toxicity, chemically reactive molecule fragments were also considered structural alarms and avoided in pharmaceuticals [5].

Structure-metabolism investigations are well-known for resolving reactive metabolite-related dangers by employing "avoidance" techniques such as structural alerts exclusion and likely termination of reactive metabolite-positive substances [6]. In one study, structural alarms were built for the generation of reactive metabolites from pharmaceuticals, and immune responses were triggered using a systematic technique including macromolecules [6, 7]. Another study employed a classifier-based technique to identify a relationship between drug substructures and protein domains. The classifier was used to extract substructures using biologically relevant chemogenomic features [8]. In a study, information about the binding location as well as the ligand's substructure was used to predict ligand-protein interaction. To extract substructure, a physical-chemical properties of binding site-based approach was used [9]. In a genome-wide screening of drug-target interactions that did not require the target protein's 3D structure, sparse canonical correspondence analysis was utilised to derive groups of chemical substructures. Following that, these substructures were utilised in the creation of a drug [10]. The RUMSSA (Relative Unified Mechanical Skill Score of Atom) technique, in addition to existing methodologies, uses X-ray crystallographic complex structure to identify potentially influential sub-structure(s) from ligand. The position of atoms in the 3D domain of a ligand-protein combination was thought to contain high selectivity information [11]. As a result, mechanical transformations of relative positioning can be employed to gather data for identifying the ligand's influential component inside the complex system of a certain protein binding pocket. As a consequence, an RUMSSA-implemented generalised approach was utilised to extract substructure from a co-crystallized protein-ligand complex's PDB dataset. The present database for cancer contains sub-structural components taken from co-crystallized ligands in PDB files.

2 Material and Method

2.1 Raw Data Collection

Protein Data Bank (PDB) files were retrieved for the express purpose of referencing human cancer sickness. These files contained at least two components, the protein and

the ligand co-crystallized with in it. Only X-ray crystallographic complex structures were used. These files contained three dimensional locations of atoms of protein and ligand.

2.2 Method to Extract Sub-structure

The RUMSSA approach was used to manually extract substructures from PDB data including co-crystallized ligands (rcsb.org). Details about the RUMSSA algorithm implemented, can be accessed from the literature 'BioRxiv 2020.02.02.931436' [11].

2.3 Organisation of Database CoSSDb

The selectivity performance of extracted substructures was carefully examined utilising experimental evidence from diverse literatures. Simple tabular data entries make up the database. The database comprises four essential pieces of information: the SMILES sub-structure, the target protein's name, class, protein and ligand ids, and facts on the target. The substructure of a ligand that may be involved in a molecular interaction with a specific protein has been extracted using an algorithm. For the hypothesising algorithm, the following assumptions were made: (i) the co-crystal structure is at its most optimised stable/stagnant state; (ii) the co-crystallized ligand has found stable interaction conformation during ligand-protein interaction; and (iv) theoretically, the dynamicity of protein and ligand will eventually achieve a combination each pegging hook is led by a single atom (referred to as the 'leader atom' in this con). Atoms of decreasing stretch follow the leader atom. With increasing distance from the leader atom, the hook's impact diminishes until it is nil. The ligand's selectivity for the target is represented by the most stretched hook atom. It was once considered that the ligand atom nearest to the protein was the most effective at attracting other ligand atoms. The 'Relative Unified Mechanical Skill Score of Atom' algorithm incorporates both the selection of the leader atom and the definition of the stretch gradient score (RUMSSA). It selectively targets a co-crystallized ligand by extracting the substructure of the ligand. The PDB file is directly processed. It's a generalised and impartial method for working with co-crystallized PDB structures. Because it was chosen from the PDB, this method ensures molecule interaction.

2.4 Algorithm Brief

Although the RUMSSA method is detailed in publication [11], but it can be summarised as follows: RUMSSA (Relative Unified Mechanical Skill Score of Atom) is a method that can determine a ligand's substructure from a PDB file comprising two molecules (Protein-Ligand) interacting. Calibration is required for each complex. As a result, the distance value should be adjusted to calibrate programme performance. The steps of the RUMSSA algorithm are as follows: The entire complex is believed to be in 3D space when processing with the full molecule. The position of each atom is traced in three dimensions. The distance between each pair of atoms is determined. A sphere is supposed to develop around each atom in this situation (the radius of the sphere is adjustable, with a default radius of 5). Each atom has an RUMSSA score of zero to

begin with. Each atom is treated in the following way: Atom pairings are chosen from the sphere's interior (i.e. have an inter-atomic distance greater than zero and a radius equal to the radius of the sphere). The atom's local RUMSSA value is updated as follows: $\left(\sum_{k=0}^j \left(\frac{1}{d_k}\right)\right)$ of Ligand within sphere around it. This step is repeated for each of the ligand atoms. The initial RUMSSA value is used to sort the atoms. The 'Leader atom' is the atom with the RUMSSA value of 'Zero.' For each Ligand-atom within a customised range, the sorted distance from the 'Leader atom' is now determined (default 5). The 'Relative RUMSSA i.e. RD_i' is now calculated in relation to the 'RUMSSA of the Leader atom' and each atom's distance (i.e. D_i) from the 'leader atom.' The RD_i vector's greatest value is utilised as a 'Threshold' for demarcating atoms starting with the leader atom. The following is the global RUMSSA value for each ligand atom:

$$R_{D_i} = \frac{1}{\left\{ \text{sort} \left(\sum_{k=0}^j \left(\frac{1}{d_k} \right) \right) \right\}_0 * e^{\left(\left(\sum_{k=0}^j (1/d_k) \right)_i * D_i \right)}$$

$$R_{D_{\text{threshold}}} = \max(R_{D_i})$$

To find the substructure that is selective for the target protein, the following conditions must be met: Atoms should have an RDThreshold Global RUMSSA value and a Local RUMSSA value that is shared by both sorted and unsorted atoms.

3 Results

A total of 7721 human cancer-related x-ray crystallography PDB files were analysed. As a result, 654 non-redundant substructures were found and added to the database. '<https://opticket49.wixsite.com/substructdb>' is the web address for the database. To access database, username & password can be used as 'omprakash'. All of these binders are linked to cancer-related protein interactions. Four different sorts of queries can be used to search databases. Search output comes in tabulated format with 'Disease,' 'SMILES,' ligand ID, PDB ID, resolution, and structural description. Substructures (1 to >1) are displayed using the SMILES format. The database website and data visualisation are depicted in Fig. 1. Chemists can be benefited much from SMILES of substructure and their annotation. Although the website's simplicity is its strength, it does have certain weakness, such as the presence of single table with multiple fields for exploratory presentation. The following queries can be used to search the database: ligand id (for example, U72, ONJ), PDB id (for example, 7NBQ, 7LT0), protein/structure name (for example, Tubulin, Estrogen), and illness name (e.g. here Cancer).

PDB co-crystallized ligand-SUB-STRUCTURE DATABASE for SELECTIVE TARGETING

(Keys for Novel drug designing & optimization for anti-cancer activity)

Enhancing Probability to Hit Target

Selective targeting is the most important aim during designing and synthesis of small molecules for future drug developments. This is a manually curated database. Sub-structures will be of key benefits for designing of potential small molecules.

DATABASE

Acknowledgment: Dr. Om Prakash expresses gratitude to Indian Council of Medical Research (ICMR) New Delhi-110029, India for providing research funding as a Research Associate Fellowship (RIS no. 2020-5197). OP is also thankful to CSIR-CMMR, Lucknow-226021, India for providing an infrastructure facility of computational biology.

SEARCH in Database

Username

Password

LOGIN

(A)

SEARCH in Database

Enter Ligand id (e.g. UT2)

Enter PDB id (e.g. 7NBQ, 7LT0)

Enter Structure Details (e.g. Tubulin, Estrogen)

Disease (e.g. Cancer)

SEARCH

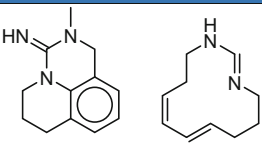
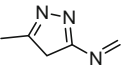
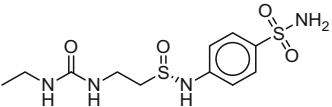
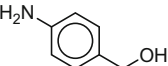
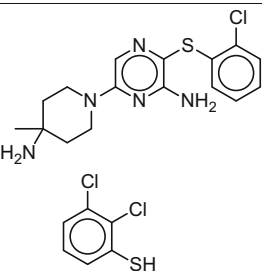
Disease	SMILES (RUMSSA Sub-Structures)	Extracted From Ligand id	Co-crystallised in PDB id	Resolution	Structure Details
Cancer	C1=NNZCOCOC3CC(C)N1C3C2C1=NNCC(C)C=CC=C(C)C	UT2	7NBQ	2.479	Co-crystal structure of Human Nicotinamide N-methyltransferase (NNMT) with the bicyclic inhibitor (4)
Cancer	CN[C@H]1CC[C@@H](C=O)C(=O)N1C2=CC=C(C)C(=O)C2	UTX	7NB1	2.691	Co-crystal structure of Human Nicotinamide N-methyltransferase (NNMT) with the bisubstrate-like inhibitor (3)
Cancer	C1=C(C)C(=O)C(=O)N1C2=CC=C(C)C(=O)C2	U7H	7NBJ	2.275	Co-crystal structure of Human Nicotinamide N-methyltransferase (NNMT) with the bisubstrate-like inhibitor (1)
Cancer	O=C(N)	ONU	7LT0	1.697	Hsp90a N-terminal inhibitor
Cancer	C1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	ONG	7LSZ	1.7	Hsp90a N-terminal inhibitor
Cancer	O=C(O)C1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	AMP	7KWM	2.3	CBP1 (28-375) L152F/V185T - AMP
Cancer	C1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	JUD	7K6P	1.4	Structure of the catalytic domain of tankyrase 1 in complex with iraparib
Cancer	SC1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	SO4	7K9Z	1.695	Structure of the catalytic domain of PARP1
Cancer	NC1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	W17	7K6E	2.47	Crystal structure of GPCR2 kinase domain gatekeeper mutant V564F in complex with covalent compound 3
Cancer	CC1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	WFD	7K6A	2.22	Crystal structure of GPCR2 kinase domain gatekeeper mutant V564F in complex with covalent compound 3
Cancer	C1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	RL4	7K6D	1.8	Estrogen Receptor Alpha Ligand Binding Domain in Complex with a Methylated Lactoferrin Derivative That Increases Receptor Resonance Time in the Nucleus of Breast Cancer Cells
Cancer	O=C1=CC=C(C=C1)C(=O)N1C2=CC=C(C)C(=O)N2	RAL	7K6A	1.78	Estrogen Receptor Alpha Ligand Binding Domain Y537S Mutant in Complex with Radioactive

(B)

Fig. 1. CoSSDb outlook. (A) Front page and (B) search result display from substructure database

Finally, the requisite sub-structure demarcation is obtained by comparing the relative commonality of mechanical transition of atomic capacity for atomic interaction. As a result, the sub-structure is determined by the global relative RUMSSA value of the atom as well as its local RUMSSA value. Here are a few samples of database output (Table 1).

Table 1. Examples of output from CoSSDb database. One or more sub-structures, from each complex, come as output from RUMSSA.

Sub-Structure	Target	PDB
	Human Nicotinamide N-methyltransferase (NNMT)	7NBQ
	FGFR2 kinase domain gatekeeper mutant V564F	7KIE
	Carbonic Anhydrase IX mimic	7K6T
	human ALDH1A1	7JWV
	Non-receptor Protein Tyrosine Phosphatase SHP2	7JVN

4 Discussion

Suggestions for substructures can be utilised to create and optimise potential leads. To execute a certain biological function, a ligand molecule interacts with one or multiple protein. For performing against a specific target protein, entire ligand molecule is not conserved. Each ligand has a structural and conformational combination in the form of one or more sub-structures, which promotes selectivity toward the target protein. To avail the hidden power sub-structure, it must be extracted from the co-crystallized ligand, so that the known sub-structures can be reused for further lead optimization. Now, the issue is that the sub-structure delineation is not well established. The RUMSSA algorithm, which was based on the mechanical behaviour of ligand atoms to define technique, has re-evolved the topic of assumption. This scheme is generalized and can be used with any ligand; because it uses a ligand itself, that is already present in a protein's 3D field domain. Here, atomic positions were redefined for expressing mechanical behaviour hypotheses for the construction of generalised schemes. In order to specify the mechanical behaviour, the reference point has been established as the leader atom. The leader atom, in theory, has the best ability to interact with protein, and the rest of the ligand atoms follow the leader's lead. As a result, the 'Relative Unified Mechanical Skill Score of Atom' was used

to establish this hypothesis theme. Atoms that followed the leader atom were thought to be a part of the potentiating substructure, while atoms that did not follow the leader atom were assumed to be excluded. The algorithm assesses the complicated molecule on both a global and local level. The global aspect looks at the entire molecule in question, whereas the local aspect looks at each individual ligand atom. This is identified as a sub-structure required for selectivity where the global and local features intersect. Because the algorithm took into account complicated molecular observation at both the global and local levels, therefore it was necessary to calibrate it in order to get reliable results. Despite the fact that RUMSSA considers all possible leader atoms, when calibrated for a single leader atom, it produces the most accurate sub-structure. The location of each atom is noticed in 3D space thanks to the algorithm's consideration of the 3D complicated PDB structure. The hidden pattern for molecular selectivity is revealed by trapped ligand atoms in the force field of protein. RUMSSA calculates the distance between each pair of atoms as a result of this process. Because each ligand atom is affected by its neighbours, a sphere with a range of area was assumed to achieve average relative influence on each atom. The technique was unbiased and generalizable to any ligand molecule because each atom was given a unified RUMSSA value. As a result, this value can be used to compare molecules. RUMSSA is a hypothetical parameter that is exclusively governed by the ligand atoms' distance from the leader atom, as optimised inside the protein force-field. Here, selective targeting was claimed because tip-of-arrow i.e. ligand side chain sub-structures were directly extracted from naturally existing complexes. Therefore, these sub-structures contain naturally defined pharmacophore for selective hitting of the target. CoSSDb is the collection of such tip-of-arrows. RUMSSA is the creator of tip-of-arrows. We processed about 7721 x-ray crystallography PDB files, and found about 654 unique substructures. These collections are ready to be attached at scaffolds for modulating their behaviour. Free accessibility of manually curated database is available at '<https://opticket49.wixsite.com/substructdb>'.

There are several previously known databases and tool, which are developed with the mind-set of cytotoxicity concerned lead optimization or toxicity related aspects in general. Benefits of CoSSDb over previous one can be observed. In a database named Swiss-Bioisosters, side chain replacement was considered for lead design. These replacement were identified through detection of matched molecular pairs and mining bioactivity data from ChEMBL database [12]. CoSSDb uses 3D structure from Protein Data Bank. In another protocol, frequency-based substructure detection protocol implemented for avoiding potential toxicity risk. It was based on sub-structures from Non-Toxic compounds [13]. CoSSDb uses RUMSSA implemented mechanical skill score of atoms for extraction of substructure. In another study, mechanisms related to specific functional groups were used for identification of sub-structures which may have capacity to trigger genotoxic or epigenetic effects [14]. CoSSDb sub-structures also show side chain along with functional groups, which has capacity to trigger modulations in signalling & expression. Manually curated substructures collections are also available, where literature contents were used for gathering information, currently available in the form of SMARTS chemical structure [15]. CoSSDb provides sub-structures in SMILES format. Which has been collected from ligand-protein co-crystallized X-ray crystallographic PDB file. Basic significance and future scope of the study can be understood as; side

chains of any drug interacts with target protein, and creates modulations in the expression or signalling extent of target protein. Therefore, significance of CoSSDb will exist at every point of ligand-protein interaction. Considering these facts, CoSSDb can be useful for target specific lead optimization by adding or removing sub-structures. Cytotoxicity of the lead molecules can also be enhanced or reduced, by using the sub-structure information from CoSSDb. Bioavailability of lead molecule can also be modulated through sub-structural modifications. Drug repurposing can also be performed by reorganising the side chains as common scaffold structures. Similarly, multiple dimensions can be touched through using sub-structures.

5 Conclusion

CoSSDb is manually curated. It has interesting substructures/bioisosteres for lead optimization and new ligand creation. They were extracted through RUMSSA algorithm, by using mechanical skill scoring of atoms. The substructures are able to directly connect to the target protein, allowing for selective targeting. This substructure database will aid in the discovery of new drugs as well as toxicological research.

Acknowledgement. Authors express gratitude to Indian council of Medical Research (ICMR) New Delhi-110029, India for providing research funding as a Research Associate Fellowship (IRIS no.: 2020-5987), as well as to CSIR-CIMAP, Lucknow-226021, India for providing an infra-structure facility for research & development in computational biology.

References

1. L. Whitesell, N. Robbins, D. S. Huang, C. A. McLellan, T. Shekhar-Guturja, E. V. LeBlanc, C. S. Nation, R. Hui, A. Hutchinson, and C. Collins, Structural basis for species-selective targeting of Hsp90 in a pathogenic fungus. *Nature communications*. 10(1), 2019, p. 1-17. DOI: <https://doi.org/10.1038/s41467-018-08248-w>
2. Takigawa, K. Tsuda, and H. Mamitsuka, Mining significant substructure pairs for interpreting polypharmacology in drug-target network. *PloS one*. 6(2), 2011, p. e16999. DOI: <https://doi.org/10.1371/journal.pone.0016999>
3. C. Valsecchi, F. Grisoni, V. Consonni, and D. Ballabio, Structural alerts for the identification of bioaccumulative compounds. *Integrated Environmental Assessment and Management*. 15(1), 2019, p. 19-28. DOI: <https://doi.org/10.1002/ieam.4085>
4. V. M. Alves, E. N. Muratov, S. J. Capuzzi, R. Politi, Y. Low, R. C. Braga, A. V. Zakharov, A. Sedykh, E. Mokshyna, and S. Farag, Alarms about structural alerts. *Green Chemistry*. 18(16), 2016, p. 4348-4360. DOI: <https://doi.org/10.1039/C6GC01492E>
5. C. Limban, D. C. Nuță, C. Chiriță, S. Negreș, A. L. Arsene, M. Goumenou, S. P. Karakitsios, A. M. Tsatsakis, and D. A. Sarigiannis, The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology reports*. 5, 2018, p. 943-953. DOI: <https://doi.org/10.1016/j.toxrep.2018.08.017>
6. S. Kalgutkar, Designing around structural alerts in drug discovery. *Journal of Medicinal Chemistry*. 63(12), 2019, p. 6276-6302. DOI: <https://doi.org/10.1021/acs.jmedchem.9b00917>
7. Claesson and A. Minidis, Systematic approach to organizing structural alerts for reactive metabolite formation from potential drugs. *Chemical Research in Toxicology*. 31(6), 2018, p. 389-411. DOI: <https://doi.org/10.1021/acs.chemrestox.8b00046>.

8. Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, and Y. Yamanishi, Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics*. 28(18), 2012, p. i487-i494. DOI: <https://doi.org/10.1093/bioinformatics/bts412>
9. Wang, J. Liu, F. Luo, Z. Deng, and Q.-N. Hu. Predicting target-ligand interactions using protein ligand-binding site and ligand substructures. in: *BMC systems biology*. Springer 2015, p. 1–10 DOI: <https://doi.org/10.1186/1752-0509-9-S1-S2>
10. Y. Yamanishi, E. Pauwels, H. Saigo, and V. Stoven, Extracting sets of chemical substructures and protein domains governing drug–target interactions. *Journal of chemical information and modeling*. 51(5), 2011, p. 1183-1194. DOI: <https://doi.org/10.1021/ci100476q>
11. O. Prakash, Algorithm for Extraction of Sub-Structure from Co-Crystallized PDB Ligand for Selective Targeting. *bioRxiv*, 2020, p. DOI: <https://doi.org/10.1101/2020.02.02.931436>
12. M. Wirth, V. Zoete, O. Michielin, and W. H. Sauer, SwissBioisostere: a database of molecular replacements for ligand design. *Nucleic acids research*. 41(D1), 2013, p. D1137-D1143. DOI: <https://doi.org/10.1093/nar/gks1059>
13. H. Yang, L. Sun, W. Li, G. Liu, and Y. Tang, Identification of nontoxic substructures: a new strategy to avoid potential toxicity risk. *Toxicological Sciences*. 165(2), 2018, p. 396-407. DOI: <https://doi.org/10.1093/toxsci/kfy146>
14. Plošnik, M. Vračko, and M. Sollner Dolenc, Mehanizmi delovanja tveganih kemijskih struktur za mutagenost in kancerogenost. *Arhiv za higijenu rada i toksikologiju*. 67(3), 2016, p. 169–182. DOI: <https://doi.org/10.1515/aiht-2016-67-2801>
15. Sushko, E. Salmina, V. A. Potemkin, G. Poda, and I. V. Tetko, ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. 2012, ACS Publications. DOI: <https://doi.org/10.1021/ci300245q>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

