Figure S1: Scheme of the validation procedure.

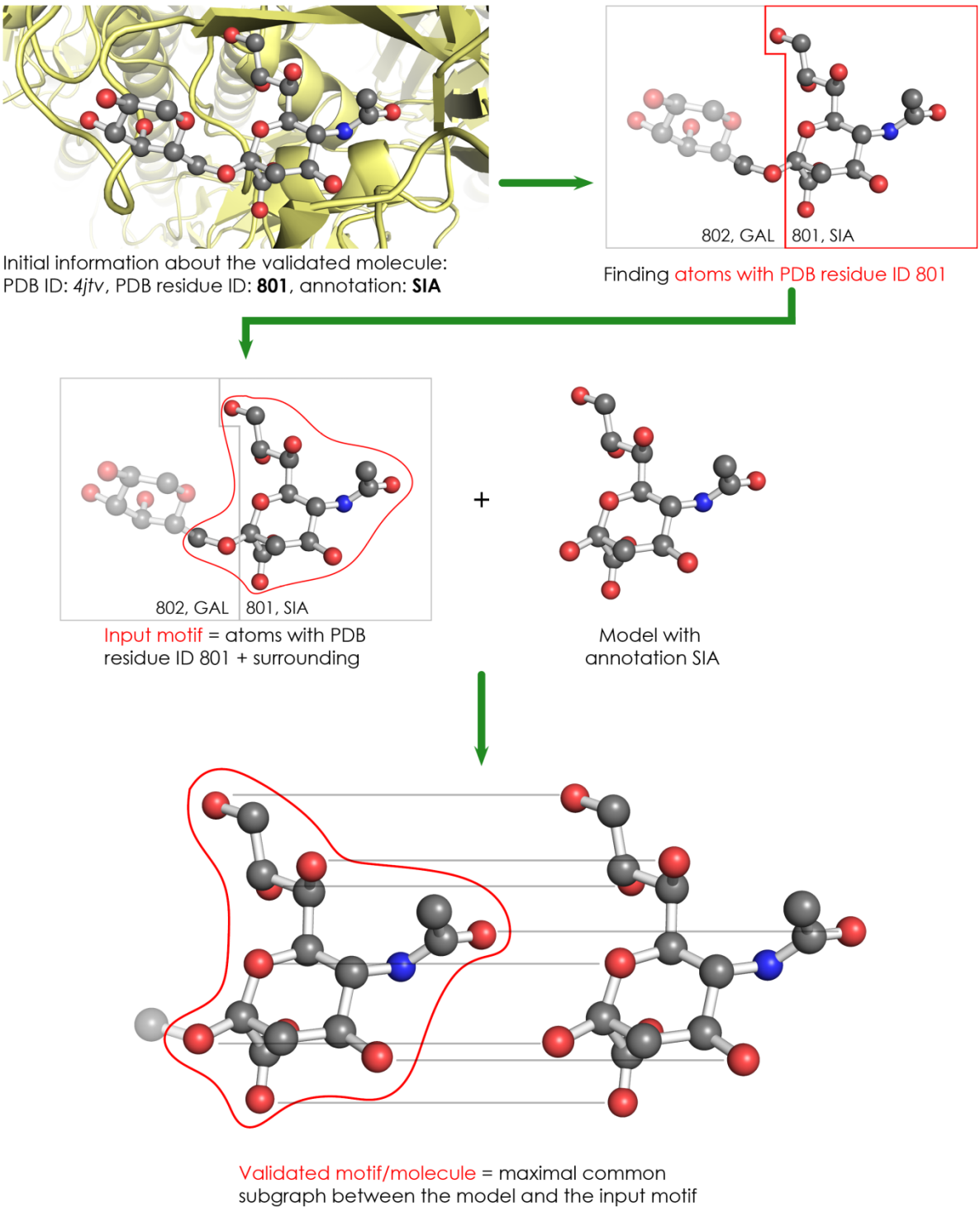

Table S1: Overview of the quality of ligands and non-standard residues in the Protein Data Bank (August 10th 2014). The evaluation involved a total of 17674 models, 238153 validated molecules, and 102364 PDB entries.

	Completeness analyses		**Chirality analyses**		**Advanced analyses**	
Wrong structures	Incomplete	8.9 %	Wrong chirality	7.9 %	-	
Issues found during individual analyses	Missing only atoms	5.9 %	Wrong C chirality	2.4 %	Atom substitution	20.7 %
	Missing rings	2.6 %	Wrong Metal chirality	1.4 %	Foreign atom	34.8 %
	Degenerate	0.5 %	Wrong High order chirality	4.3 %	Different atom naming	38.2 %
			Wrong Planar chirality	1.1 %	Alternate conformations	2.4 %
No significant issues	**Complete**	91.1 %	**Complete + Correct chirality**	83.0 %	**Complete + Correct chirality + no Atom substitutions + no Foreign atoms**	48.0%
			Complete + Correct chirality (tolerant)	88.3 %	**Complete + Correct chirality (tolerant) + no Atom substitutions + no Foreign atoms**	53.2%

Table S2: Summarization of validation results for individual case studies and their comparison with the validation results for all ligands and non-standard residues in the Protein Data Bank (from August 10th 2014).

	All molecules	Polycyclic	Carbo-hydrates	Mannose derivates	Organo-metals	Experi-mental drugs	Approved drugs
Number of PDB entries analyzed	102364	3568	8752	1534	5216	15307	958
Number of validated molecules	238153	6804	57302	6341	22600	37450	1934
Number of models used as reference	17674	1370	913	53	331	3399	185
Incomplete	8.9%	6.7%	5.9%	3.5%	18.0%	6.1%	3.2%
Missing only atoms	5.9%	3.1%	4.2%	3.0%	6.0%	5.0%	0.9%
Missing rings	2.6%	3.0%	1.5%	0.1%	10.7%	0.6%	2.0%
Degenerate	0.5%	0.6%	0.2%	0.4%	1.4%	0.5%	0.3%
Wrong chirality	7.9%	5.5%	4.0%	7.6%	16.5%	2.1%	2.8%
Wrong C chirality	2.4%	3.5%	4.0%	7.4%	2.5%	1.7%	2.8%
Wrong Metal chirality	1.4%	0.0%	0.0%	0.0%	14.3%	0.0%	0.0%
Wrong High order chirality	4.3%	1.9%	0.0%	0.2%	0.0%	0.4%	0.0%
Wrong Planar chirality	1.1%	0.0%	0.0%	0.0%	10.5%	0.1%	0.8%
Complete	**91.1%**	**93.3%**	**94.1%**	**96.5%**	**82.1%**	**93.9%**	**96.8%**
Complete + Correct chirality	**83.0%**	**87.6%**	**90.1%**	**88.9%**	**64.3%**	**91.8%**	**93.9%**
Complete + Correct chirality (tolerant)	**88.3%**	**89.9%**	**90.1%**	**89.1%**	**75.0%**	**92.2%**	**94.7%**

Legend: The color code refers to the relative difference between the results of each case study and the PDB-wide average for all ligands and non-standard residues. Specifically:

> 2 times better
> 30% better
> 30% worse
> 2 times worse

Table S3: Sources of molecules, which are visualized in Figure 1. All models were taken from wwPDB CCD, according to their annotation (3-letter code). The label "generated" refers to structures which were not extracted from real PDB entries, but were generated in order to illustrate the types of validation analyses included in ValidatorDB in a clear and unified manner.

Completeness analyses	**Description**	**a)**	**b)**	**c)**	**d)**
	Source	model	generated	generated	generated
Chirality analyses	**Description**	**e) correct**	**f) correct**	**g) correct**	**h) correct**
	Source	model	model	model	model
	Description	**e) wrong**	**f) wrong**	**g) wrong**	**h) wrong**
	Source	generated	2p7e: 5 A	1arx: 345 A	4a2u: 280 C
Advanced analyses	**Description**	**a)**	**i)**	**j)**	**k)**
	Source	model	generated	generated	generated

Algorithm limitations

The algorithm behind ValidatorDB has the following limitations:

- It is necessary to ensure that the model serving as reference during validation is indeed correct. This limitation is overcome by using high-quality structures from wwPDB CCD.
- The superimposition phase might not identify the optimal matching between the atoms of the model and those of the validated molecule if their 3D structures are too different. Specifically, the molecule may appear severely fragmented if some critical atoms are missing or misplaced (i.e., the length of the bond connecting that atom to the rest of the structure differs by over 3 standard deviations from the general PDB average expected for that bond type). In this case, the molecules are generally marked as degenerate. This limitation applies to no more than 0.3% of validated molecules across the entire Protein Data Bank.
- The superimposition phase might not identify the optimal matching if the validated molecule contains very complicated scaffolds like cages. In such cases, the molecules may incorrectly be marked as degenerate (e.g., heptamolibdate XBO). This limitation applies to no more than 0.5% of validated molecules across the entire Protein Data Bank, and is generally seen in organometallic cages.
- Some molecules are counted as alternate conformers even if they are not marked by the standard alternate location identifiers in the PDB file. Such situations arise when two molecules with the same annotation (3-letter code) but different residue identifiers lie closer than 0.65 Å from each other. In this case, only the molecule with the lower residue identifier is validated. Alternate conformers, either explicitly marked as such in the PDB file or not, are not validated. They add up to approximately 2.5% of the molecules relevant for validation across the entire Protein Data Bank.

All limitations, except for the first one, cause the particular molecule to be marked by an explicit processing warning in all validation reports.

Case study details: Wiki page containing a link to validation results and a list with validated annotations (3-letter codes) for each case study.

http://webchem.ncbr.muni.cz/Wiki/ValidatorDB:CaseStudies