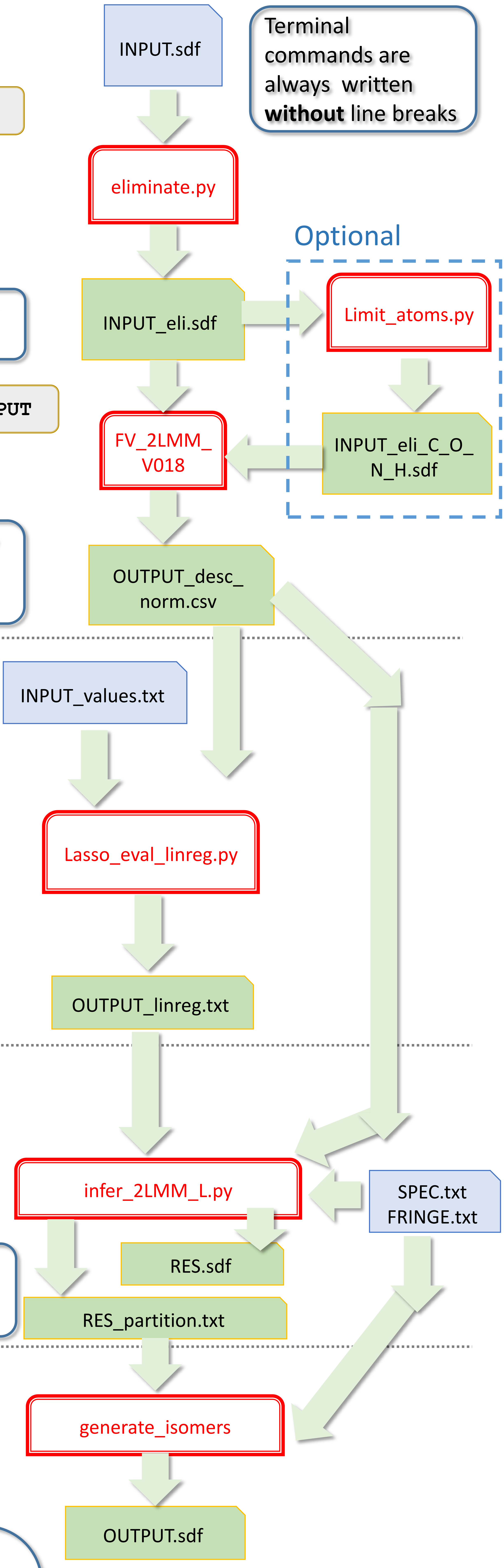
Files and Flow of Data in 2LMM-LLR

Used symbols and conventions

- Files provided by the User
- Files produced by the programs
- Executable program
- `Terminal command`



Terminal commands are always written **without** line breaks

Module 1 (Calculating Feature Vectors)

```
python eliminate.py INPUT.sdf
```

Limit_atoms.py

The extension "_eli.sdf" is automatically added to the input file name.

```
./FV_2LMM_V018 INPUT_eli.sdf OUTPUT
```

The name of the output file is decided by the user and part of the command arguments.

Module 2 (Training an hyperplane)

```
python lasso_eval_linreg.py
OUTPUT_desc_norm.csv
INPUT_values.txt
OUTPUT 0.01 (0.01 is the
penalty constant for Lasso)
```

- The suffix "_linreg.txt" is added automatically.
- The filename "OUTPUT" is decided by the user.

Module 3 (MILP)

```
python
  infer_2LMM_L.py
  OUTPUT 5 42 SPEC.txt
   FRINGE.txt RES.sdf
```

"5 and 42" are lower and upper bounds, resp., for a target value given by the user

Module 4 (Generating Graph Structures)

```
./generate_isomers
  RES.sdf 2 10000 5 10 10000 2
  OUTPUT.sdf RES_partition.txt
  FRINGE.txt
```

-"2" time limit for each stage of program
- "10000" is an upper bound on feature vectors stored in memory
- "5" is the number of sample graphs stored for each feature vector
- "10" upper bound on time for enumeration of paths
- "10000" upper bound on the number of total paths stored during the computation
- "2" upper bound on the number of output graphs

"OUTPUT.sdf" is the final result of this model.
In this example, it has up to 2 chemical graphs with structure specification given in the file "RES.sdf", that have target value between "5 and 42".