SimulationsPlus

MIDD+22

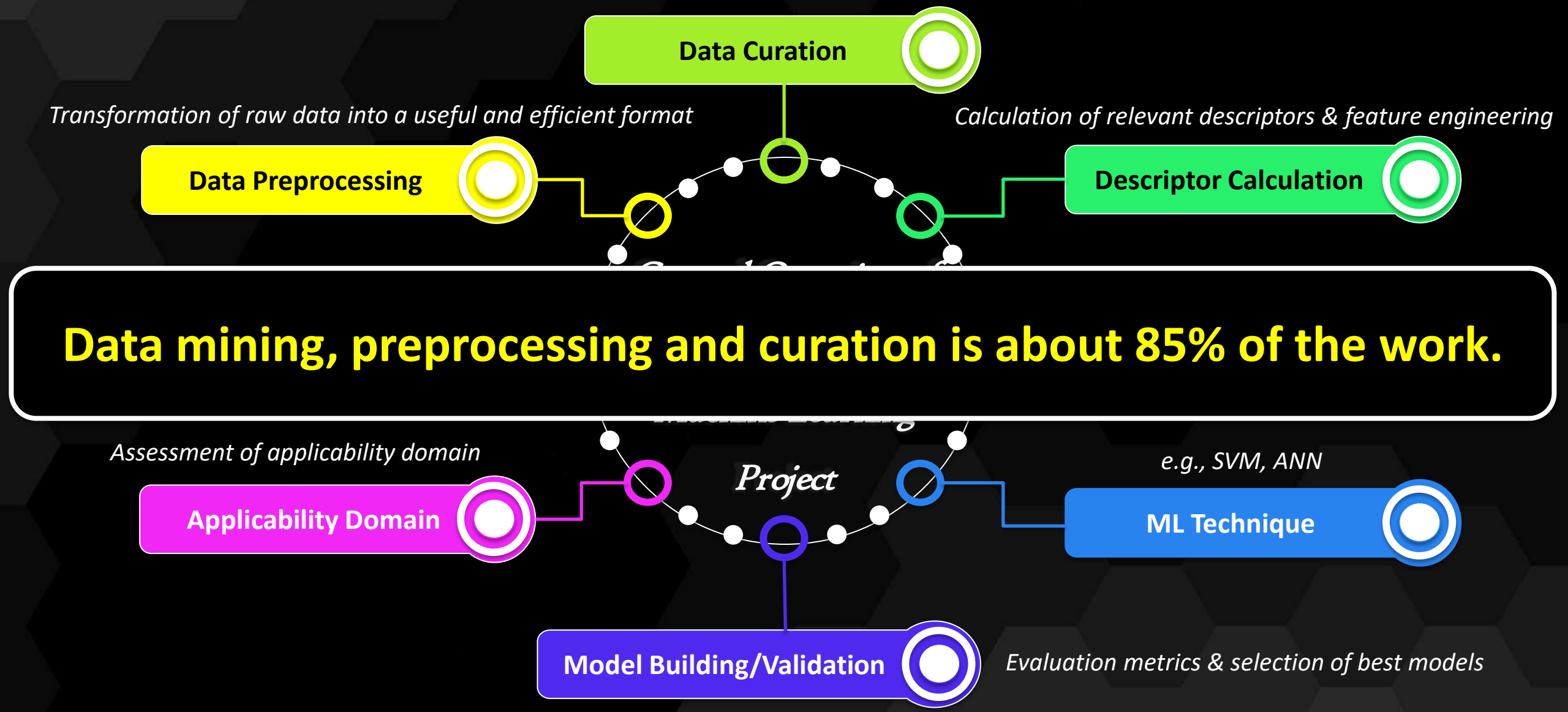
Model Informed Drug Development

How We Leverage Automation for Data Mining, Preprocessing & Curation

Phyo Phyo Kyaw Zin

[• REC] Please note: this presentation, including questions from the audience, is being recorded and may be made available.

Ensure dataset is high quality; containing biologically relevant and correct data for biological endpoints of interest



When Do We Use Automation?

- When there are tedious, repetitive, exhausting & time-consuming tasks for humans
- When we can improve quality of work, efficiency, and reduce human error
- When there is a clear, well-defined set of rules to address common problems in the pipeline

Example Cases of Automation in CYP450 site of metabolism (SoM) Project

Web Scraping for Drugs

- Extraction of drug names with possible CYP information from FDA website in an efficient & fast way
- Scan documents for CYP information to assess priority

Structural Extraction & Verification

- Retrieval of chemical structures based on names from multiple databases using APIs
- Structural preprocessing & comparison of structures

Merging Datasets

Merging different datasets based on SMILES, handling conflicted entries, removing duplicates, & flagging suspicious CYP enzymes that should be inspected manually

OUTPUT

1	Drug Name	Active Ingredient	Approval Date	FDA-approved use	links	keywords_detected	priority
16	Avycaz	ceftazidime-aviba	2/25/2015	To treat adults wit	http://www.acces	cyp, 1a2, 2a6, 2b6, 2c8, 2c9, 2c19,	high
17	Axumin	fluciclovine F 18	5/27/2016	A new diagnostic i	http://www.accessdata.fda.gov/scripts/cder/daf/inde		low
18	Ayvakit	avapritinib	1/9/2020	To treat adults wit	http://www.acces	cyp, 1a2, 2b6, 2c8, 2c9, 2c19, 2d6,	high
19	Balversa	erdafitinib	4/12/2019	To treat adult pati	http://www.acces	cyp, 2c9, 3a4	high
20	Barhemsys	amisulpride	2/26/2020	To help prevent na	http://www.acces	cyp, 1a2, 2a6, 2b6, 2c8, 2c9, 2c19,	high
21	Baxdela	delafloxacin	6/19/2017	To treat patients v	http://www.accessdata.fda.gov/scripts/cder/daf/inde		low
22	benznidazole	benznidazole	8/29/2017	To treat children a	http://www.acces	cyp, 1a2, 2b6, 3a4	high
23	Bevyxxa	betrixaban	6/23/2017	For the prophylaxi	http://www.acces	cyp, 1a2, 2b6, 2c9, 2c19, 2d6, 3a4	high
24	Biktarvy	bictegravir, embit	2/7/2018	To treat infection	http://www.acces	cyp	high
25	Braftovi	encorafenib	6/27/2018	To treat unresecta	http://www.acces	cyp, 1a2, 2b6, 2c8, 2c9, 2c19, 2d6,	high
26	Bridion	sugammadex	12/15/2015	To reverse effects	http://www.accessdata.fda.gov/scripts/cder/daf/inde		low
34	Cerianna	fluoroestradiol F1	5/20/2020	Diagnostic imagin	http://www.accessdata.fda.gov/scripts/cder/daf/inde		low
35	Cholbam	cholic acid	3/17/2015	To treat pediatric	http://www.acces	cyp	high
36	Copiktra	duvelisib	9/24/2018	To treat relapsed	http://www.acces	cyp, 3a4	high
37	Corlanor	ivabradine	4/15/2015	To reduce hospita	http://www.acces	cyp, 3a4	high
38	Cotellic	cobimetinib	11/10/2015	To be used in com	http://www.acces	cyp, 1a2, 2b6, 2c8, 2c9, 2c19, 2d6,	high
39	Cresembacapsulei	isavuconazonium	3/6/2015	To treat adults wit	http://www.accessdata.fda.gov/scripts/cder/daf/inde		low
40	Daklinza	daclatasvir	7/24/2015	To treat chronic h	http://www.acces	cyp, 1a2, 2b6, 2c8, 2c9, 2c19, 2d6,	high
41	Daurismo	glasdegib	11/21/2018	To treat newly-dia	http://www.acces	cyp, 1a2, 2b6, 2c8, 2c9, 2c19, 2d6,	high
42	Dayvigo	lemborexant	12/20/2019	To treat insomnia	http://www.acces	cyp, 1a2, 2a6, 2b6, 2c8, 2c9, 2c19,	high
43	Defitelio	defibrotide sodiur	3/30/2016	To treat adults a	http://www.acces	cyp, 1a2, 2b6, 2c8, 2c9, 2c19, 2d6,	high

Showing 1 to 1 of 1 entries

nation Extracted

Drug Info

Text Scan

Assess Priority

This automation pipeline efficiently mines hundreds of drugs and prioritizes them for SoM project in minutes.

Model Informed Drug Development +

SimulationsPlus
Cognigen | DILIsym Services | Lixoft

Example Cases of Automation in CYP450 site of metabolism (SoM) Project

Web Scraping for Drugs

- Extraction of drug names with possible CYP information from FDA website in an efficient & fast way
- Scan documents for CYP information to assess priority

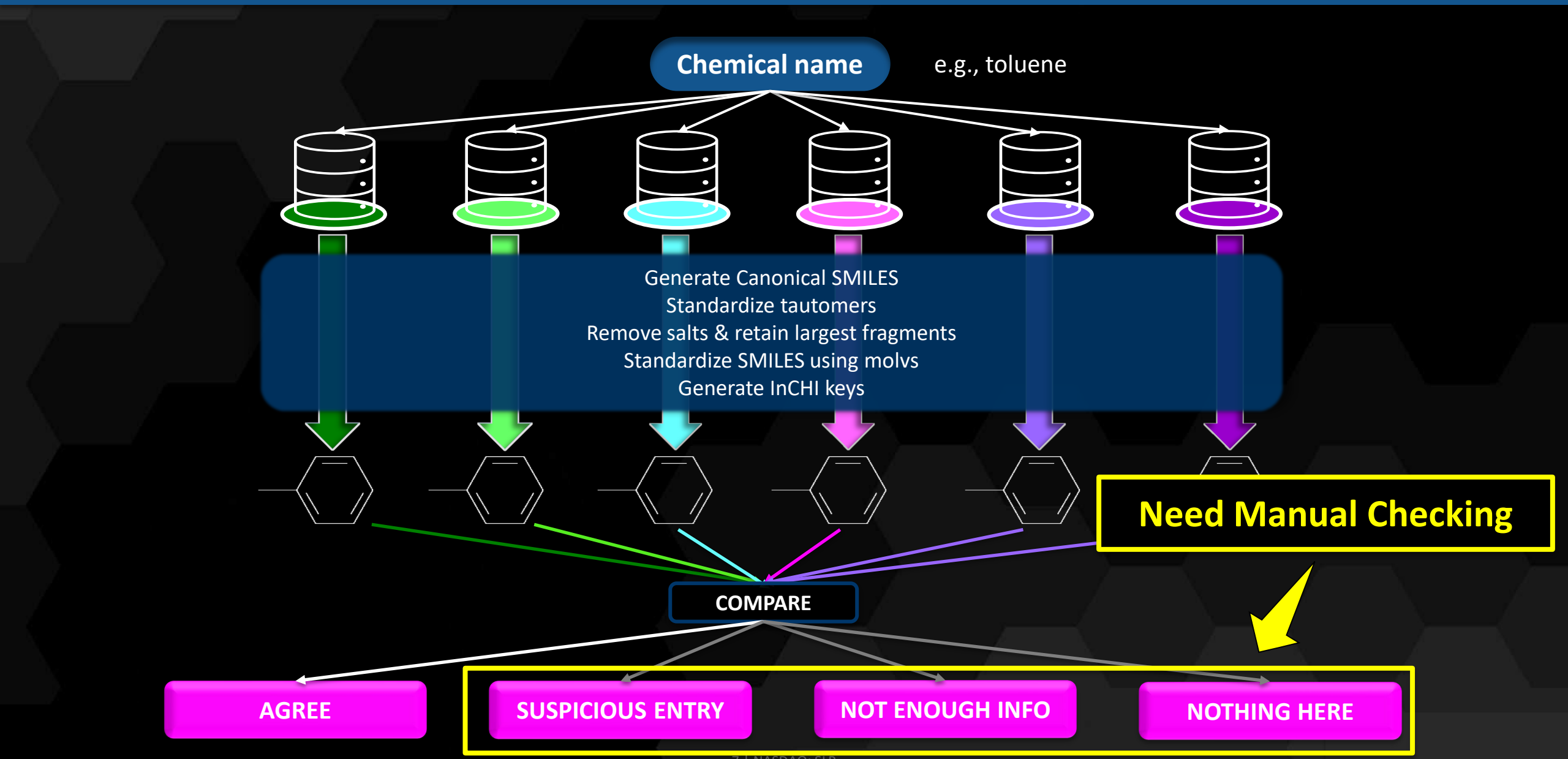
Structural Extraction & Verification

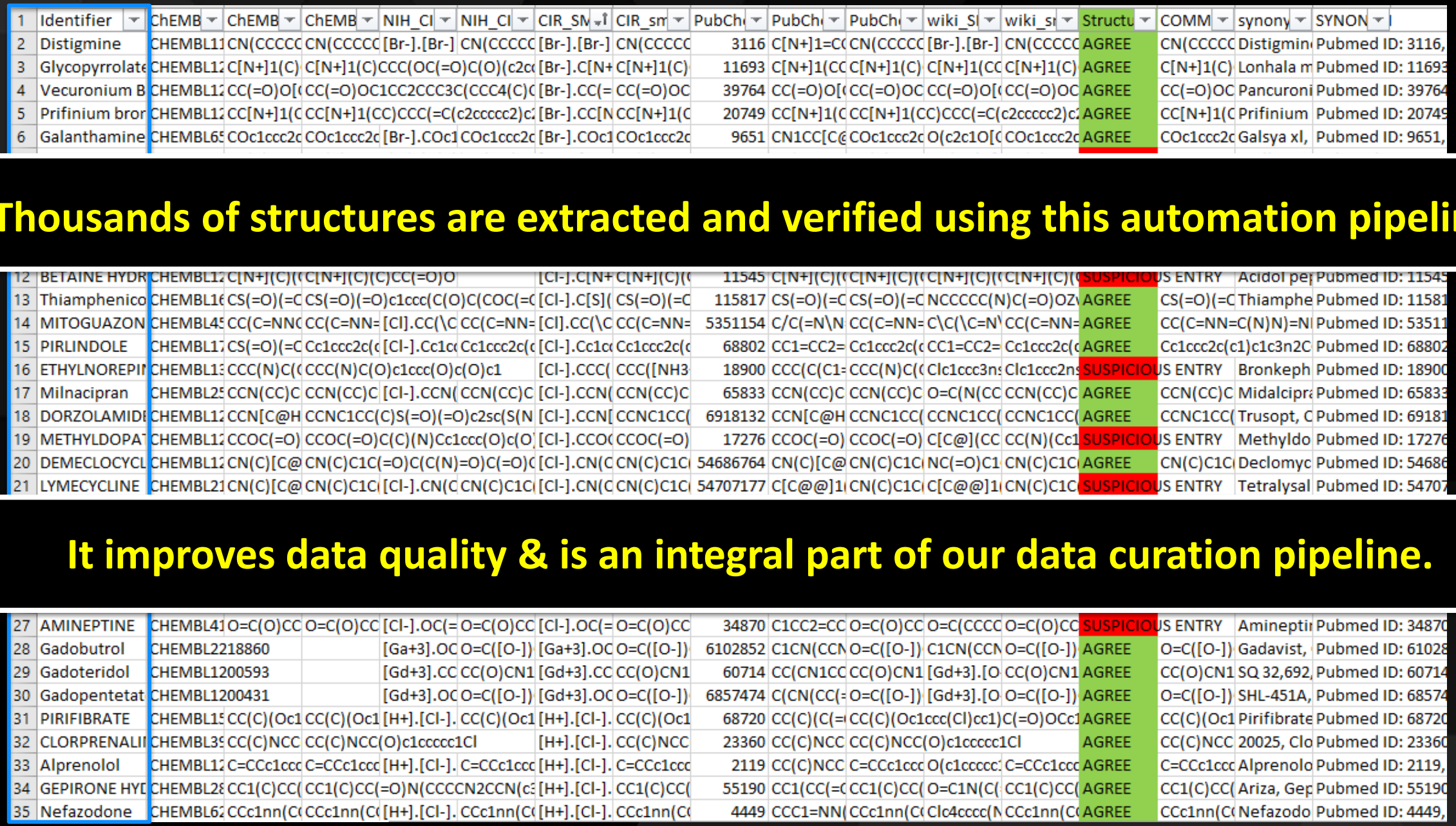
- Retrieval of chemical structures based on names from multiple databases using APIs
- Structural preprocessing & comparison of structures

Merging Datasets

Merging different datasets based on SMILES, handling conflicted entries, removing duplicates, & flagging suspicious CYP enzymes that should be inspected manually

Automation for Structural Retrieval & Verification





	Identifier	ChEMB	ChEMB	ChEMB	NIH_Cl	NIH_Cl	CIR_SM	CIR_sm	PubCh	PubCh	PubCh	wiki_Sl	wiki_sr	Structu	COMM	synony	SYNON
2	Distigmine	CHEMBL11	CN(CCCCC	CN(CCCCC	[Br-].[Br-]	CN(CCCCC	[Br-].[Br-]	CN(CCCCC	3116	C[N+]1=C(	CN(CCCCC	[Br-].[Br-]	CN(CCCCC	AGREE	CN(CCCCC	Distigmin	Pubmed ID: 3116,
3	Glycopyrrolate	CHEMBL12	C[N+]1(C)	C[N+]1(C)CCC(OC(=O)C(O)(c2c	[Br-].C[N+	C[N+]1(C)	11693	C[N+]1(CC	C[N+]1(C)	C[N+]1(CC	C[N+]1(C)	AGREE	C[N+]1(C)	Lonhala m	Pubmed ID: 11693
4	Vecuronium B	CHEMBL12	CC(=O)O[(	CC(=O)OC1CC2CCC3C(CCC4(C)C	[Br-].CC(=	CC(=O)OC	39764	CC(=O)O[(	CC(=O)OC	CC(=O)O[(	CC(=O)OC	AGREE	CC(=O)OC	Pancuroni	Pubmed ID: 39764
5	Prifinium bror	CHEMBL12	CC[N+]1(C	CC[N+]1(CC)CCC(=C(c2ccccc2)c2	[Br-].CC[N	CC[N+]1(C	20749	CC[N+]1(C	CC[N+]1(CC)CCC(=C(c2ccccc2)c2	AGREE	CC[N+]1(C	Prifinium	Pubmed ID: 20749
6	Galanthamine	CHEMBL65	COc1ccc2c	COc1ccc2c	[Br-].COc1	COc1ccc2c	[Br-].COc1	COc1ccc2c	9651	CN1CC[C@	COc1ccc2c	O(c2c1O[(	COc1ccc2c	AGREE	COc1ccc2c	Galsya xl,	Pubmed ID: 9651,
12	BETAINE HYDR	CHEMBL12	C[N+](C)((	C[N+](C)(C)CC(=O)O		[Cl-].C[N+	C[N+](C)((	11545	C[N+](C)((	C[N+](C)((	C[N+](C)((	C[N+](C)((	SUSPICIOUS ENTRY		Acidol pep	Pubmed ID: 11545
13	Thiamphenico	CHEMBL16	CS(=O)(=C	CS(=O)(=O)c1ccc(C(O)C(COC(=(	[Cl-].C[S](	CS(=O)(=C	115817	CS(=O)(=C	CS(=O)(=C	NCCCCC(N)C(=O)OZ	AGREE	CS(=O)(=C	Thiamphe	Pubmed ID: 11581
14	MITOGUAZON	CHEMBL45	CC(C=NNC	CC(C=NN=	[Cl].CC(\C	CC(C=NN=	[Cl].CC(\C	CC(C=NN=	5351154	C/C(=N\N	CC(C=NN=	C\C(\C=N\	CC(C=NN=	AGREE	CC(C=NN=C(N)N)=N	Pubmed ID: 53511
15	PIRLINDOLE	CHEMBL17	CS(=O)(=C	Cc1ccc2c(c	[Cl-].Cc1c	Cc1ccc2c(c	[Cl-].Cc1c	Cc1ccc2c(c	68802	CC1=CC2=	Cc1ccc2c(c	CC1=CC2=	Cc1ccc2c(c	AGREE	Cc1ccc2c(c1)c1c3n2C	Pubmed ID: 68802
16	ETHYLNOREPI	CHEMBL13	CCC(N)C((	CCC(N)C(O)c1ccc(O)c(O)c1	[Cl-].CCC(	CCC([NH3	18900	CCC(C(C1=	CCC(N)C((	Clc1ccc3ns	Clc1ccc2ns	SUSPICIOUS ENTRY		Bronkeph	Pubmed ID: 18900
17	Milnacipran	CHEMBL25	CCN(CC)C	CCN(CC)C	[Cl-].CCN(	CCN(CC)C	[Cl-].CCN(	CCN(CC)C	65833	CCN(CC)C	CCN(CC)C	O=C(N(CC	CCN(CC)C	AGREE	CCN(CC)C	Midalcipra	Pubmed ID: 65833
18	DORZOLAMIDE	CHEMBL12	CCN[C@H	CCNC1CC(C)S(=O)(=O)c2sc(S(N	[Cl-].CCN[	CCNC1CC(	6918132	CCN[C@H	CCNC1CC(	CCNC1CC(	CCNC1CC(	AGREE	CCNC1CC(	Trusopt, C	Pubmed ID: 69181
19	METHYLDOPA	CHEMBL12	CCOC(=O)	CCOC(=O)C(C)(N)Cc1ccc(O)c(O	[Cl-].CCO(	CCOC(=O)	17276	CCOC(=O)	CCOC(=O)	C[C@](CC	CC(N)(Cc1	SUSPICIOUS ENTRY		Methyldo	Pubmed ID: 17276
20	DEMECLOCYCL	CHEMBL12	CN(C)[C@	CN(C)C1C(=O)C(C(N)=O)C(=O)C	[Cl-].CN(C	CN(C)C1C(	54686764	CN(C)[C@	CN(C)C1C(	NC(=O)C1	CN(C)C1C(	AGREE	CN(C)C1C(	Declomyc	Pubmed ID: 54686
21	LYMECYCLINE	CHEMBL21	CN(C)[C@	CN(C)C1C(	[Cl-].CN(C	CN(C)C1C(	[Cl-].CN(C	CN(C)C1C(	54707177	C[C@@]1(	CN(C)C1C(	C[C@@]1(	CN(C)C1C(	SUSPICIOUS ENTRY		Tetralysal	Pubmed ID: 54707
27	AMINEPTINE	CHEMBL41	O=C(O)CC	O=C(O)CC	[Cl-].OC(=	O=C(O)CC	[Cl-].OC(=	O=C(O)CC	34870	C1CC2=CC	O=C(O)CC	O=C(CCCC	O=C(O)CC	SUSPICIOUS ENTRY		Amineptir	Pubmed ID: 34870
28	Gadobutrol	CHEMBL2218860			[Ga+3].OC	O=C([O-])	[Ga+3].OC	O=C([O-])	6102852	C1CN(CCN	O=C([O-])	C1CN(CCN	O=C([O-])	AGREE	O=C([O-])	Gadavist,	Pubmed ID: 61028
29	Gadoteridol	CHEMBL1200593			[Gd+3].CC	CC(O)CN1	[Gd+3].CC	CC(O)CN1	60714	CC(CN1CC	CC(O)CN1	[Gd+3].[O	CC(O)CN1	AGREE	CC(O)CN1	SQ 32,692,	Pubmed ID: 60714
30	Gadopentetat	CHEMBL1200431			[Gd+3].OC	O=C([O-])	[Gd+3].OC	O=C([O-])	6857474	C(CN(CC(=	O=C([O-])	[Gd+3].[O	O=C([O-])	AGREE	O=C([O-])	SHL-451A,	Pubmed ID: 68574
31	PIRIFIBRATE	CHEMBL15	CC(C)(Oc1	CC(C)(Oc1	[H+].[Cl-].	CC(C)(Oc1	[H+].[Cl-].	CC(C)(Oc1	68720	CC(C)(C(=	CC(C)(Oc1ccc(Cl)cc1)C(=O)OCc1	AGREE	CC(C)(Oc1	Pirifibrate	Pubmed ID: 68720
32	CLORPRENALI	CHEMBL39	CC(C)NCC	CC(C)NCC(O)c1ccccc1Cl		[H+].[Cl-].	CC(C)NCC	23360	CC(C)NCC	CC(C)NCC(O)c1ccccc1Cl		AGREE	CC(C)NCC	20025, Clo	Pubmed ID: 23360
33	Alprenolol	CHEMBL12	C=CCc1ccc	C=CCc1ccc	[H+].[Cl-].	C=CCc1ccc	[H+].[Cl-].	C=CCc1ccc	2119	CC(C)NCC	C=CCc1ccc	O(c1ccccc	C=CCc1ccc	AGREE	C=CCc1ccc	Alprenolo	Pubmed ID: 2119,
34	GEPIRONE HYD	CHEMBL28	CC1(C)CC(	CC1(C)CC(=O)N(CCCCN2CCN(c3	[H+].[Cl-].	CC1(C)CC(	55190	CC1(CC(=C	CC1(C)CC(	O=C1N(C(	CC1(C)CC(	AGREE	CC1(C)CC(	Ariza, Gep	Pubmed ID: 55190
35	Nefazodone	CHEMBL62	CCc1nn(C(	CCc1nn(C(	[H+].[Cl-].	CCc1nn(C(	[H+].[Cl-].	CCc1nn(C(	4449	CCC1=NN(	CCc1nn(C(	Clc4cccc(N	CCc1nn(C(	AGREE	CCc1nn(C(	Nefazodo	Pubmed ID: 4449,

Thousands of structures are extracted and verified using this automation pipeline.

It improves data quality & is an integral part of our data curation pipeline.

Example Cases of Automation in CYP450 site of metabolism (SoM) Project

Web Scraping for Drugs

- Extraction of drug names with possible CYP information from FDA website in an efficient & fast way
- Scan documents for CYP information to assess priority

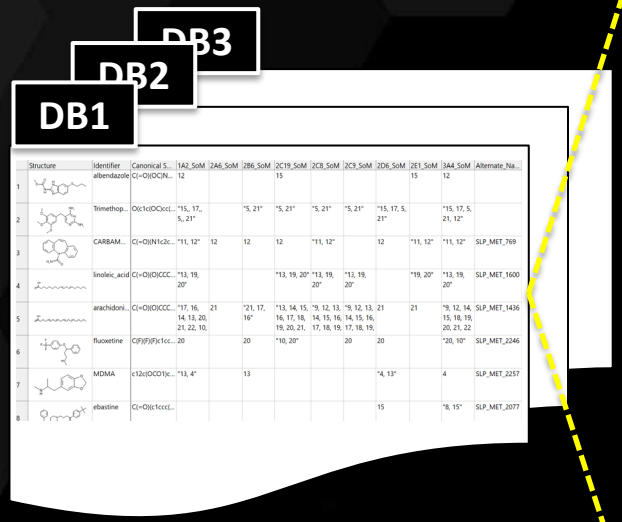
Structural Extraction & Verification

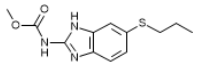
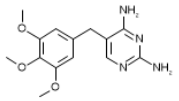
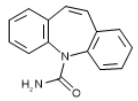
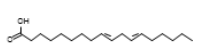
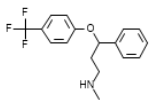
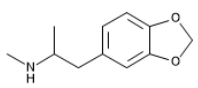
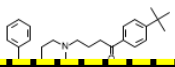
- Retrieval of chemical structures based on names from multiple databases using APIs
- Structural preprocessing & comparison of structures

Merging Datasets

Merging different datasets based on SMILES, handling conflicted entries, removing duplicates, & flagging suspicious CYP enzymes that should be inspected manually

Automation in Merging Datasets



	Structure	Identifier	Canonical S...	1A2_SoM	2A6_SoM	2B6_SoM	2C19_SoM	2C8_SoM	2C9_SoM	2D6_SoM	2E1_SoM	3A4_SoM	Alternate_Na...
1		albendazole	C(=O)(OC)N...	12			15				15	12	
2		Trimethop...	O(c1c(OC)cc(...	"15,, 17,, 5,, 21"		"5, 21"	"5, 21"	"5, 21"	"5, 21"	"15, 17, 5, 21"		"15, 17, 5, 21, 12"	
3		CARBAM...	C(=O)(N1c2c...	"11, 12"	12	12	12	"11, 12"		12	"11, 12"	"11, 12"	SLP_MET_769
4		linoleic_acid	C(=O)(O)CCC...	"13, 19, 20"			"13, 19, 20"	"13, 19, 20"	"13, 19, 20"		"19, 20"	"13, 19, 20"	SLP_MET_1600
5		arachidoni...	C(=O)(O)CCC...	"17, 16, 14, 13, 20, 21, 22, 10,	21	"21, 17, 16"	"13, 14, 15, 16, 17, 18, 19, 20, 21,	"9, 12, 13, 14, 15, 16, 17, 18, 19,	"9, 12, 13, 14, 15, 16, 17, 18, 19,	21	21	"9, 12, 14, 15, 18, 19, 20, 21, 22	SLP_MET_1436
6		fluoxetine	C(F)(F)(F)c1cc...	20		20	"10, 20"		20	20		"20, 10"	SLP_MET_2246
7		MDMA	c12c(OCO1)c...	"13, 4"		13				"4, 13"		4	SLP_MET_2257
8		ebastine	C(=O)(c1ccc(...							15		"8, 15"	SLP_MET_2077

Automation in Merging Datasets

When handling duplicate structures …

The tasks involved in the data-merging process are redundant, time-consuming and highly susceptible to human mistakes.

Identifier	Canonical SMILES	1A2_SoM	2A6_SoM	2B6_SoM	2C19_SoM	2C8_SoM	2C9_SoM	2D6_SoM	2E1_SoM	3A4_SoM	Alternate Name	Reference	Dataset
Compound A	O=C1c2c(ncn2C)N(C(=O)N1CCCCC(O)C)C	7	7	7			35 34 7	7		7 29		10.1093/bioinformatics/btw617	DB1

Identifier	Canonical SMILES	1A2_SoM	2A6_SoM	2B6_SoM	2C19_SoM	2C8_SoM	2C9_SoM	2D6_SoM	2E1_SoM	3A4_SoM	Alternate Name	Reference	Dataset
Compound II	O=C1c2c(ncn2C)N(C(=O)N1CCCCC(O)C)C	7	8	1			34 29 37 7	7		34 29 37 7		10.1073/pnas.2016778118	DB3

This automation pipeline merges multiple datasets in a few seconds while following consistent rules in the process.

Compound A	O=C1c2c(ncn2C)N(C(=O)N1CCCCC(O)C)C	7	7, 8	1, 7			7, 29, 34, 35, 37	7		7, 29, 34, 37	Compound II	formatics/btw617, 10.1073/pnas.2016778118	DB1, DB3	2B6_SoM, 2C9_SoM, 3A4_SoM

Conclusion

- Three example use cases of automation in CYP450 site of metabolism (SoM) project were shown to highlight the efficiency, usefulness and importance of automation.

- Automation is often used in data-mining, preprocessing and curation.
  - To reduce tedious, repetitive, exhausting & time-consuming tasks for humans
  - To diminish human mistakes/manual errors
  - To enhance data quality
  - To improve efficiency by lowering time, cost and effort

  (esp. when handling thousands/millions of data)

Acknowledgement

- Pankaj Daga
- Michael Lawless
- Bob Clark
- David Miller
- Marvin Waldman
- Robert Fraczkiewicz
- Viera Lukacova
- Eric Jamois
- John Dibella

SimulationsPlus

MIDD+22

Model Informed Drug Development

Q&A

Questions & Answers

phyophyo.zin@simulations-plus.com

www.drzinph.com