

Computer-assisted structure identification (CASI) for high-throughput identification of small molecules by GC \times GC–HRAM–TOFMS

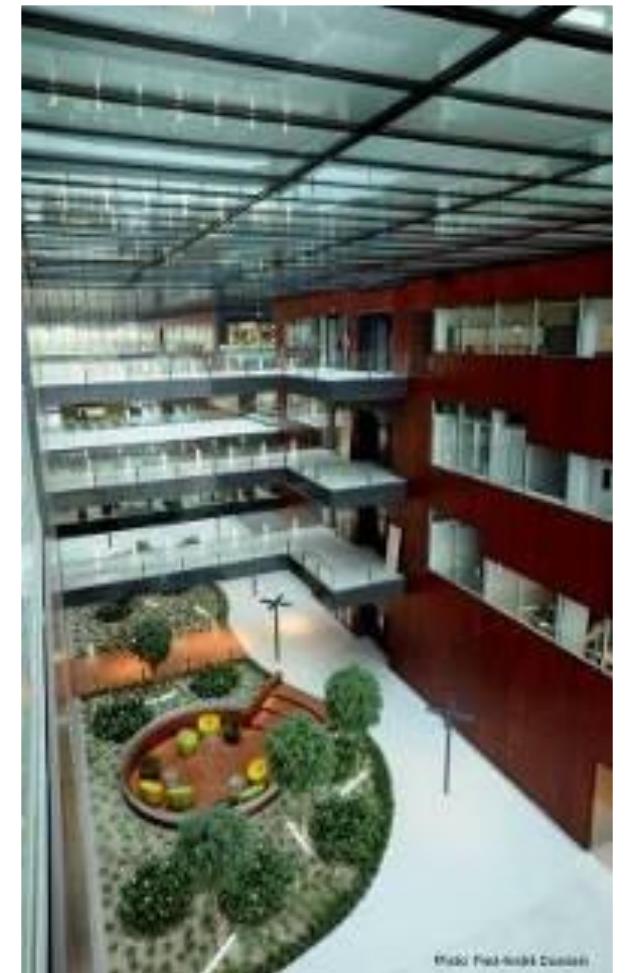
Arno Knorr, Elyette Martin, Martin Almstetter, Antonio Castellon, Pavel Pospisil, Mark Bentley, Catherine Goujon
PMI R&D, Philip Morris Products S.A., CH-2000 Neuchâtel, Switzerland



2020 CORESTA ONLINE Congress
12 October – 12 November 2020

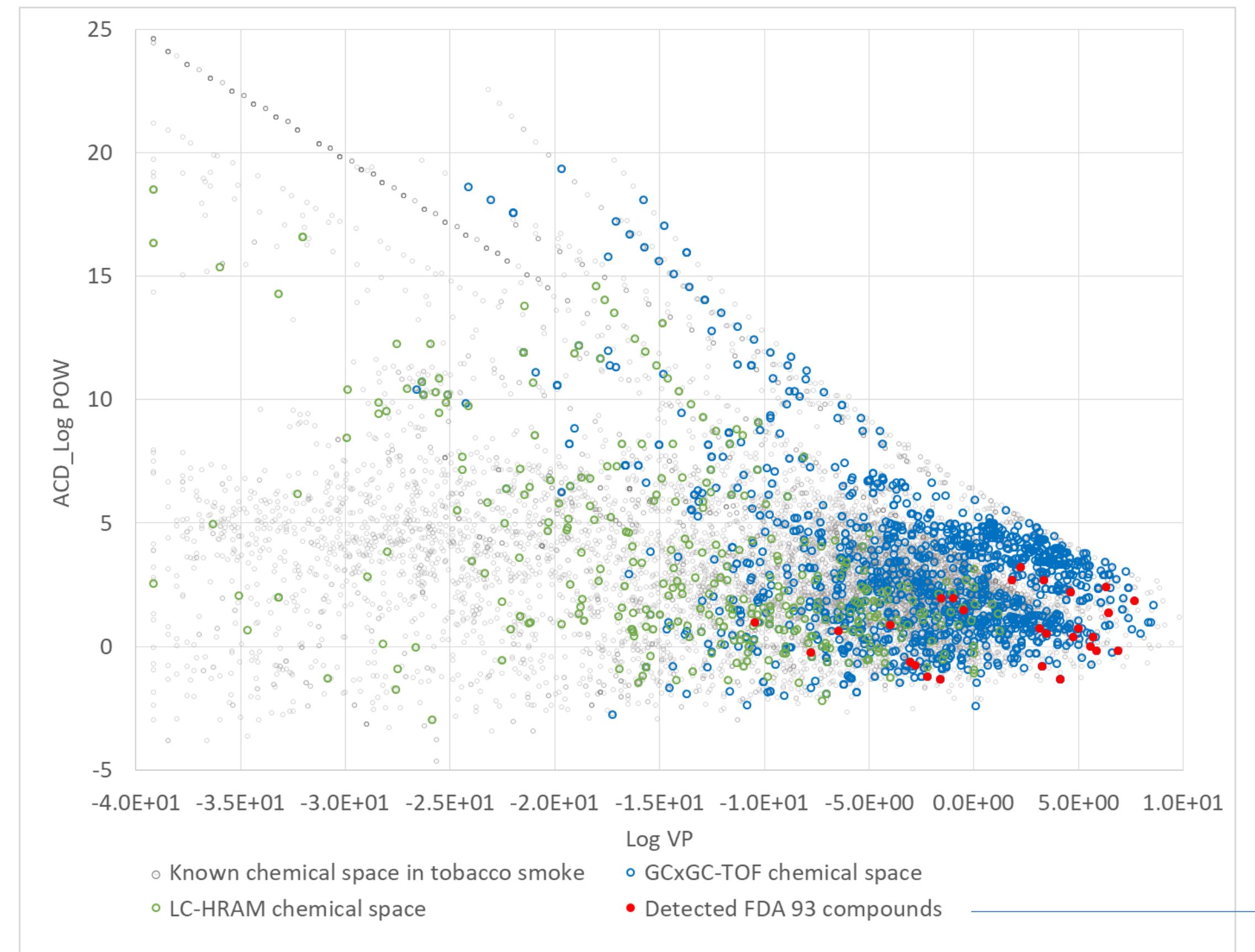
Non-Targeted Screening Talks by PMI R&D @CORESTA

- ST 17** **The dos and don'ts of non-targeted screening by LC–HRAM-MS for chemical characterization of smoke-free products**
1682 WACHSMUTH C.; ARNDT D.; BUCHHOLZ C.; BENTLEY M.; GOUJON C.
Philip Morris Products S.A., PMI R&D, Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland
- ST 18** **Computer-assisted structure identification (CASI) for high-throughput identification of small molecules by GC \times GC–HRAM-TOFMS**
1683 KNORR A.; ALMSTETTER M.; MARTIN E.; CASTELLON A.; POSPISIL P.; BENTLEY M.; GOUJON C.
Philip Morris Products S.A., PMI R&D, Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland
- ST 19** **Non-targeted chemical characterization of complex matrices by nominal- and high-resolution accurate-mass GC \times GC–TOFMS**
1703 ALMSTETTER M.; KNORR A.; RHOUMA M.; MARTIN E.; CASTELLON A.; POSPISIL P.; BENTLEY M.; GOUJON C.
Philip Morris Products S.A., PMI R&D, Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland
- ST 21** **Untargeted chemical characterization of the aerosol generated by a heated tobacco product**
1799 BENTLEY M.; ALMSTETTER M.; ARNDT D.; KNORR A.; MARTIN E.; POSPISIL P.; MAEDER S.
Philip Morris Products S.A., PMI R&D, Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland



Aerosol chemistry — Comprehensive chemical characterization

Non-Targeted Screening (NTS) of THS2.2 aerosol and 3R4F cigarette smoke — Chemical space Complementary characters of GCxGC-TOFMS and LC-HRAM-MS data



Coverage of chemical space by non-targeted GC \times GC-TOFMS (3 methods) and LC-HRAM-MS (4 methods) in the context of the known chemical space of tobacco plant & smoke (more than 6,000 for cigarette smoke)

FDA 93 compounds: A small subset of the chemical space covered by our NTS methods

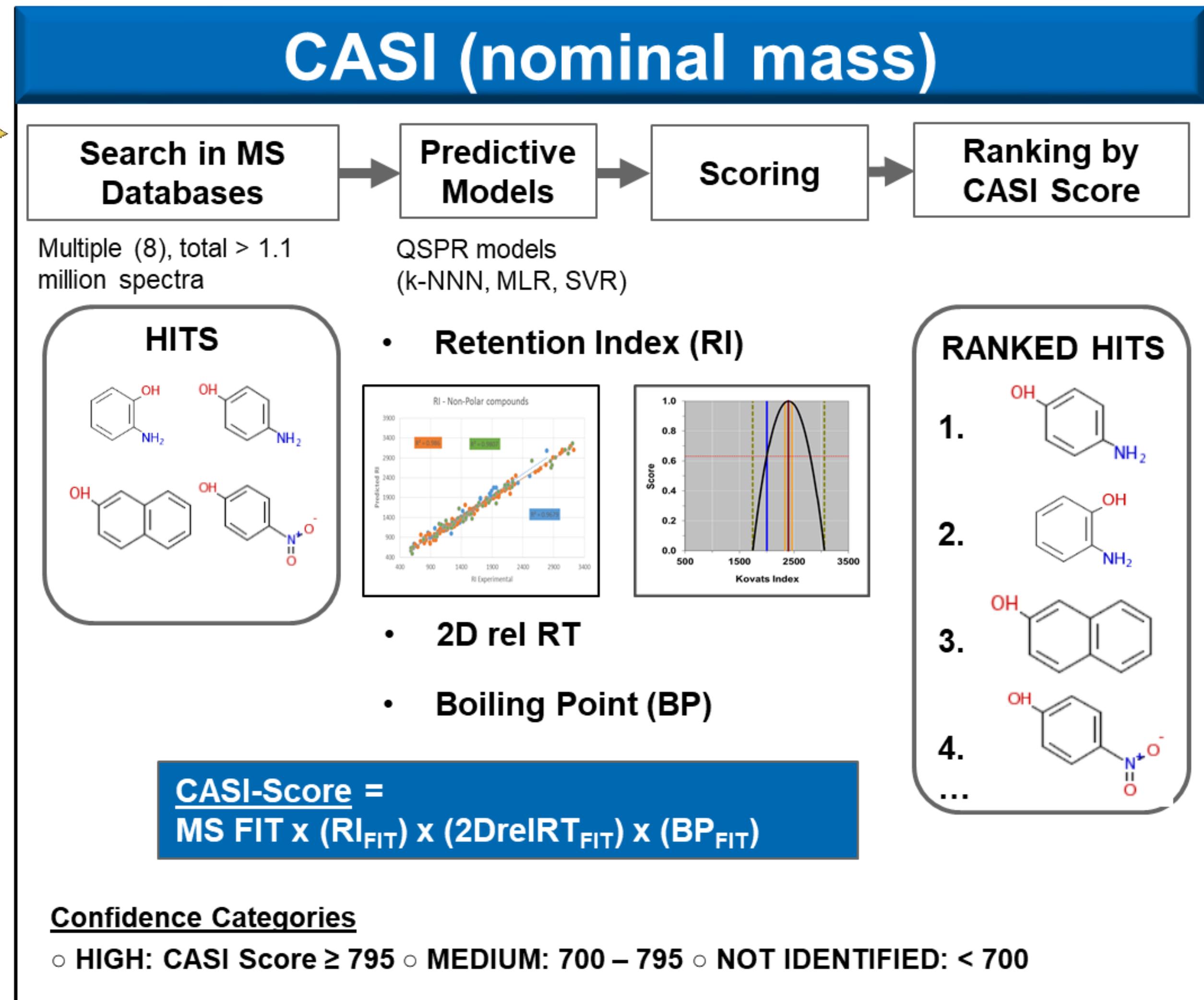
- Coverage by overlapping individual methods/platforms to avoid potential gaps in the chemical space covered

ACD_Log POW: logarithm of octanol/ water partition coefficient values,
Log VP: logarithm of vapor pressure both calculated using ACD/Labs Percepta suite software.

Knorr, A., Almstetter, M. et al., *Analytical Chemistry*, 2019
Arndt, A., Wachsmuth, C. et al., *Rapid Communications in Mass Spectrometry*, 2019

High-throughput structure identification (GC — Nominal mass)

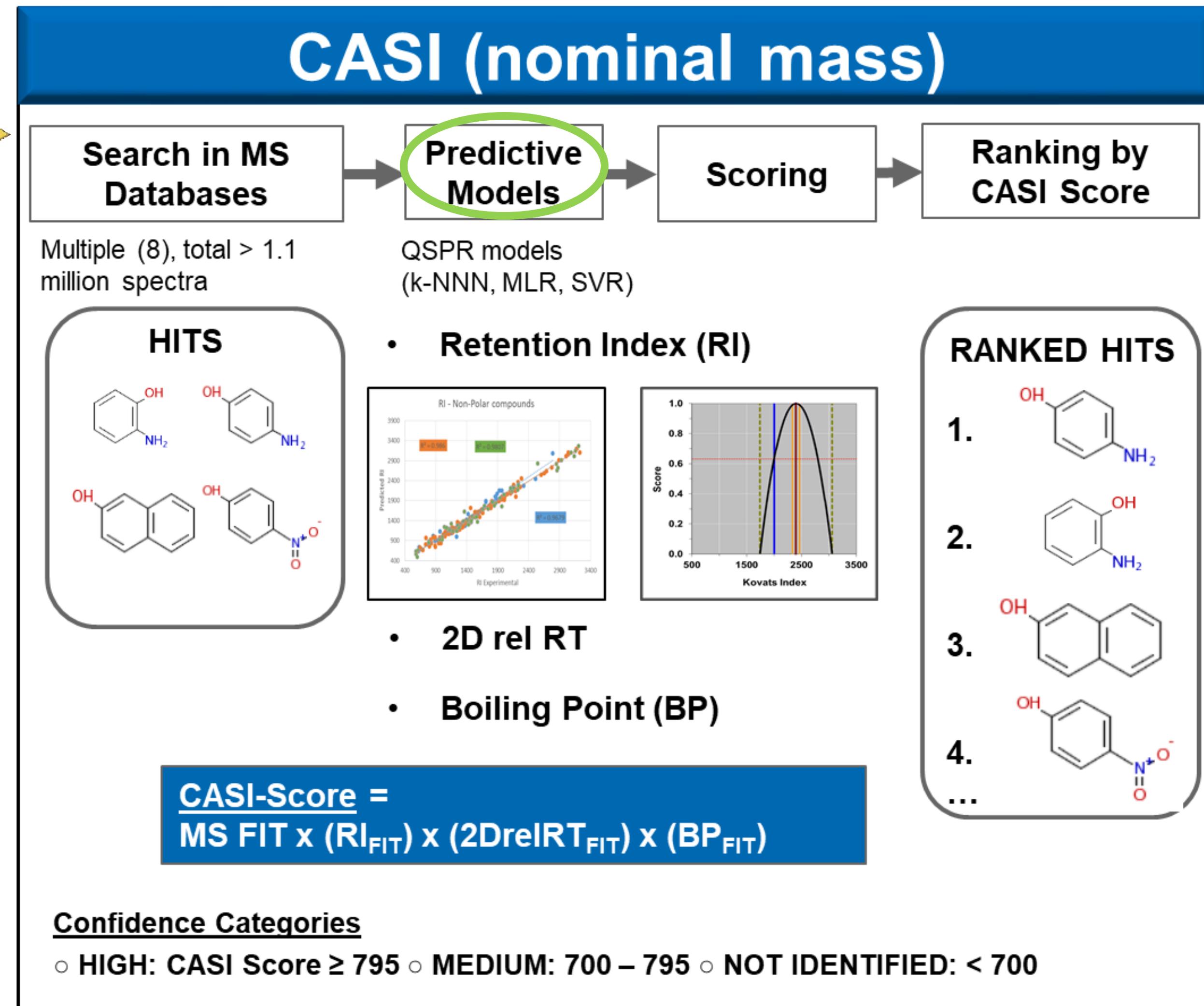
Computer-assisted structure identification (CASI) for GCxGC-TOFMS



Knorr, A., et al., Analytical Chemistry 2013, 85, 11216
<https://www.ncbi.nlm.nih.gov/pubmed/24160557>

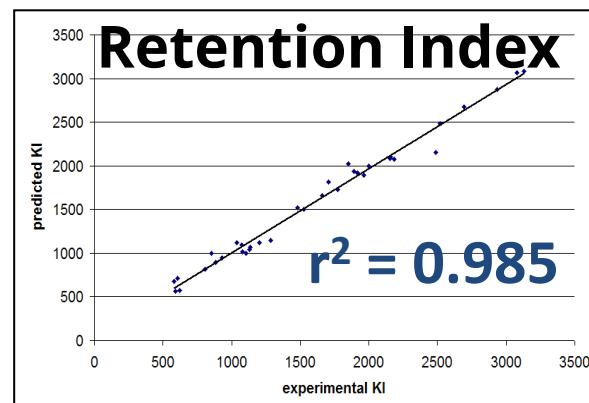
High-throughput structure identification (GC — Nominal mass)

Computer-assisted structure identification (CASI) for GCxGC-TOFMS

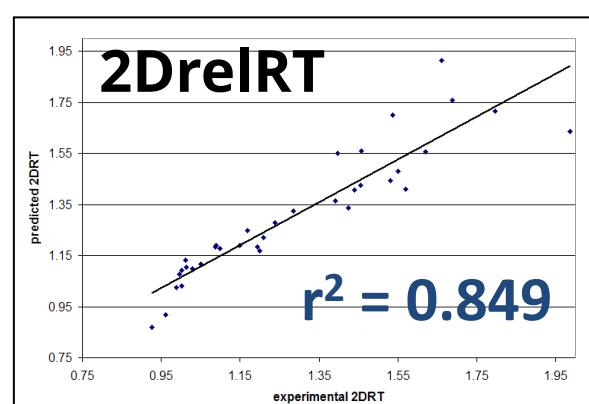


RI 1st dimension retention index
 2DrelRT 2nd dimension relative retention time
 GA Genetic algorithms
 Molecular Descriptors Numerical values associated with the chemical constitution for correlation of chemical structure with various physical properties, chemical react., or biol. activity

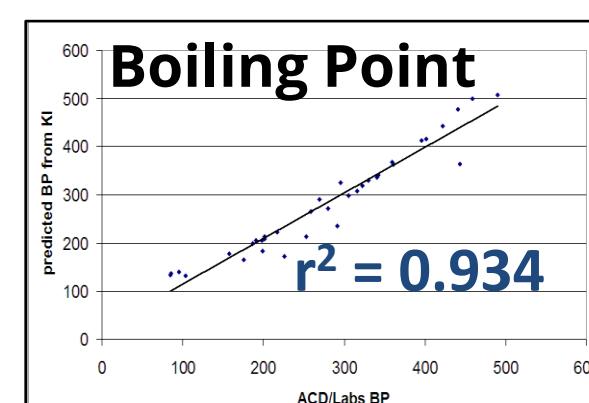
Predictive QSPR Models



GA – Linear Regression,
15 Molecular Descriptors

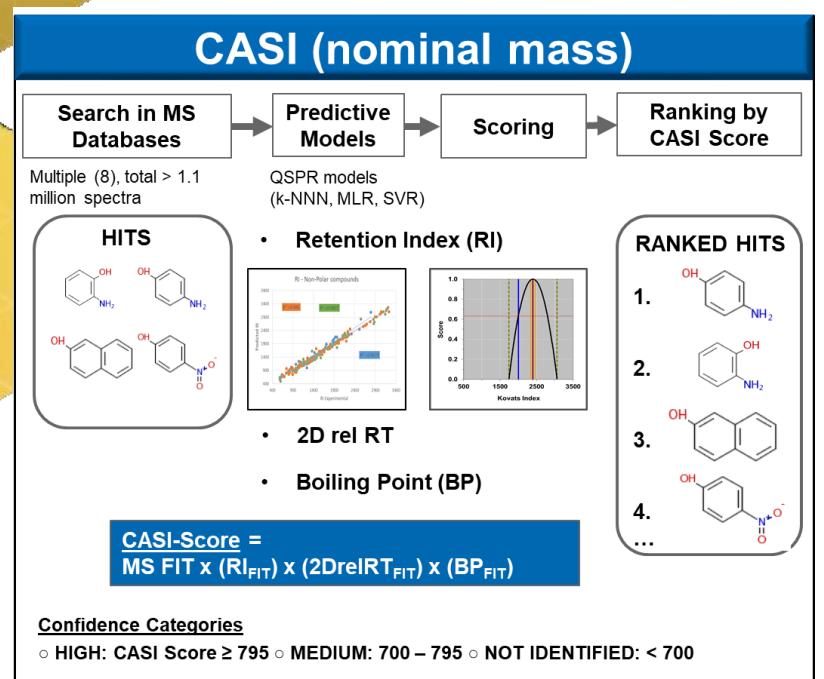


GA – Support Vector Regression,
8 Molecular Descriptors

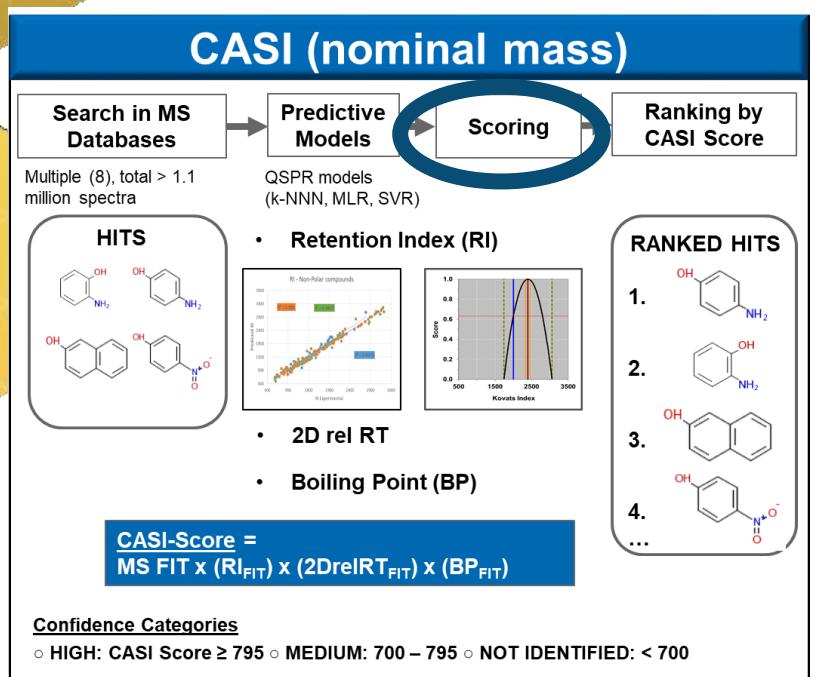


Linear Regression: BP calc. by
ACD/PhysChem vs. BP calc. by KI

High-throughput structure identification (GC — Nominal mass)

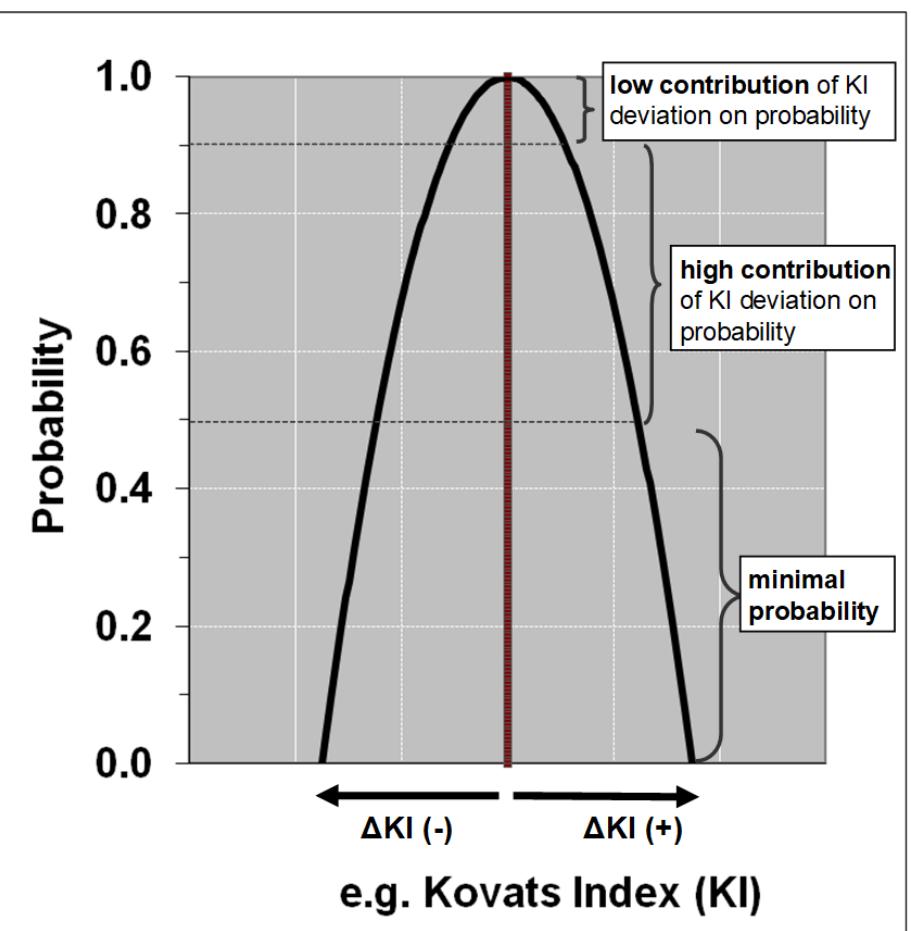


High-throughput structure identification (GC — Nominal mass)

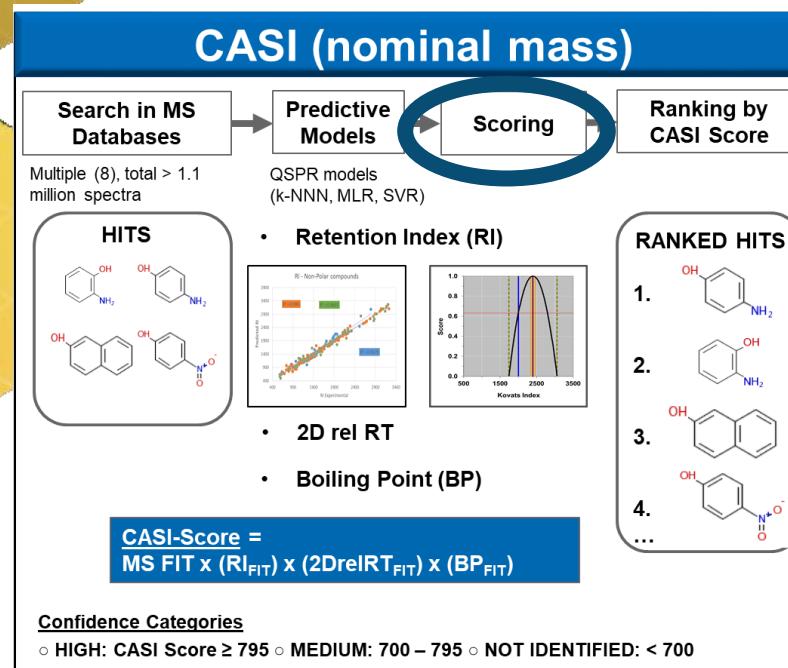


Score functions

- Parabolic function for each analytical property

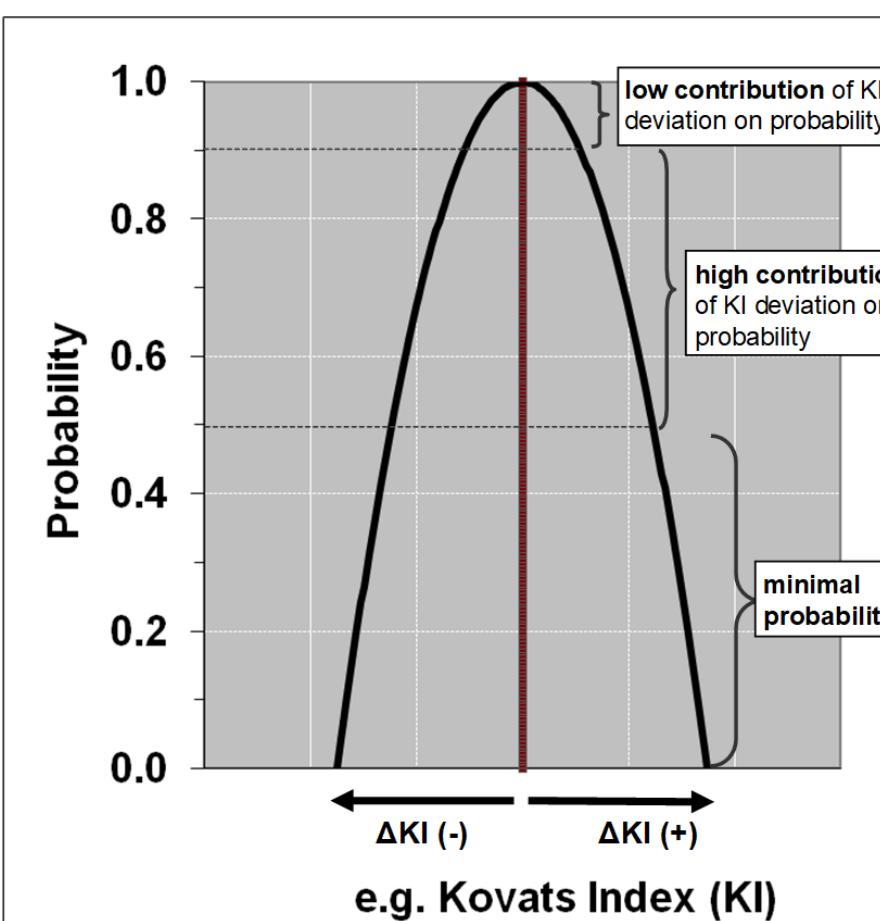


High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

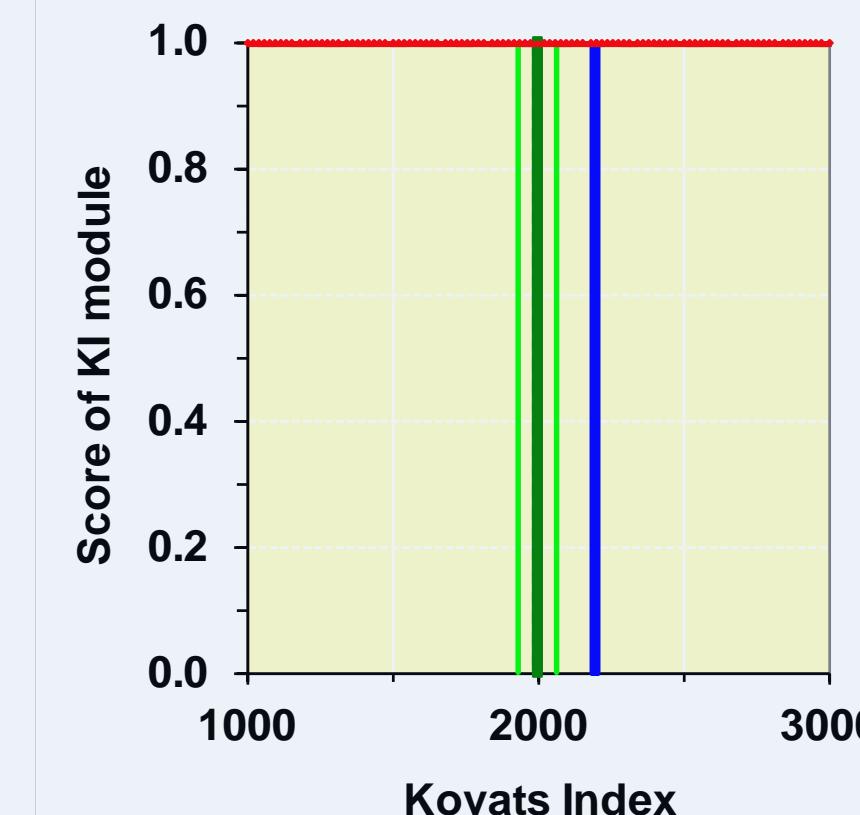
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	1000.0
Score	1.000

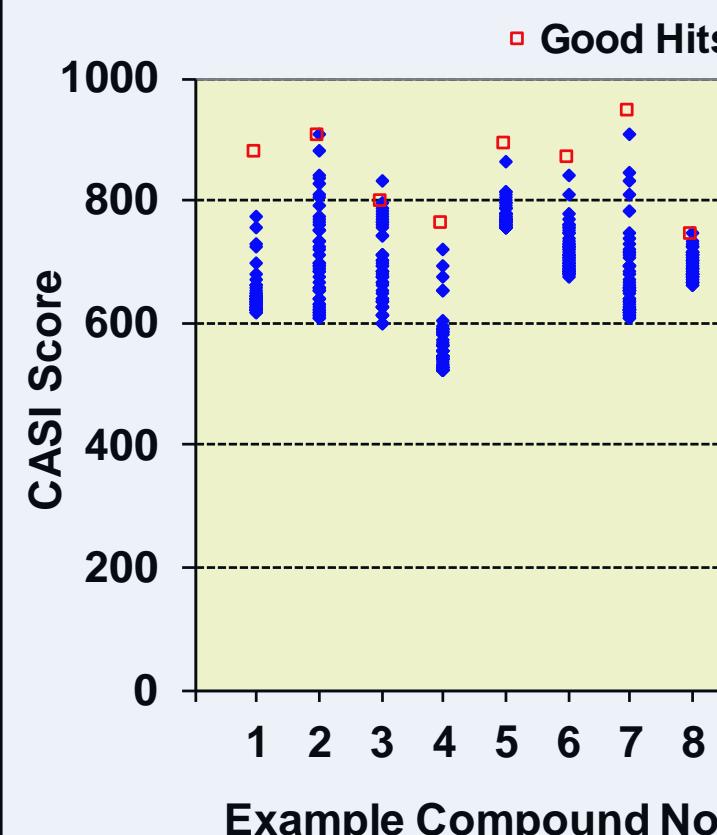
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

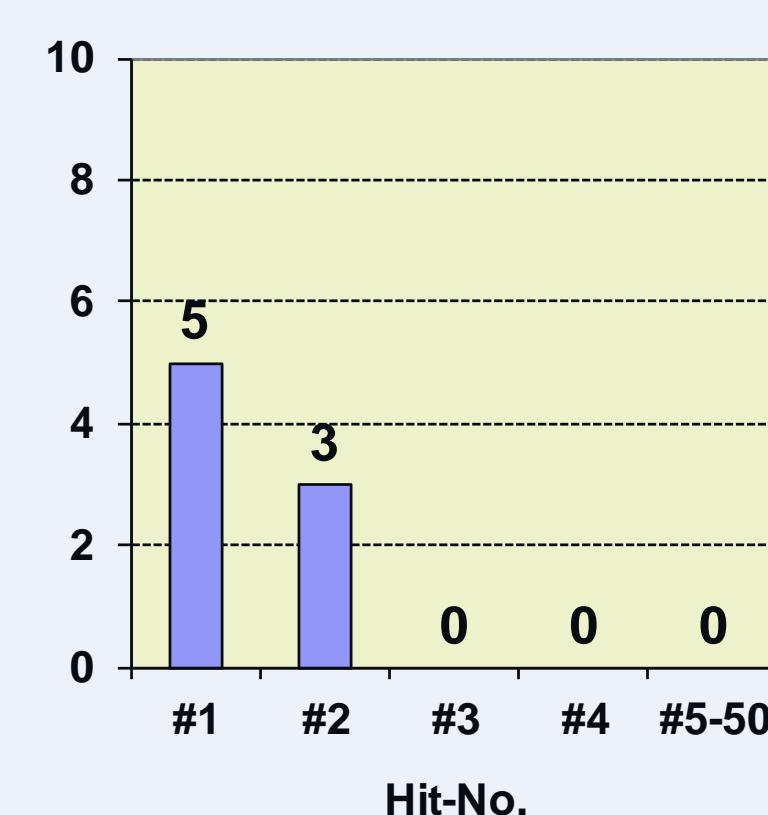
Visualization of curve fitting



Score by MS Similarity and Predicted KI



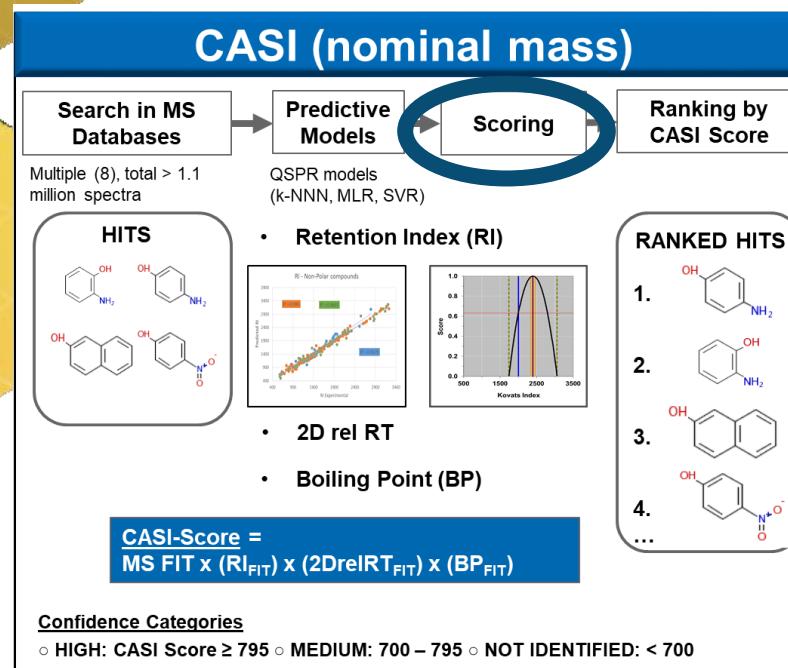
Hit ranking of correct structures



Score model optimization

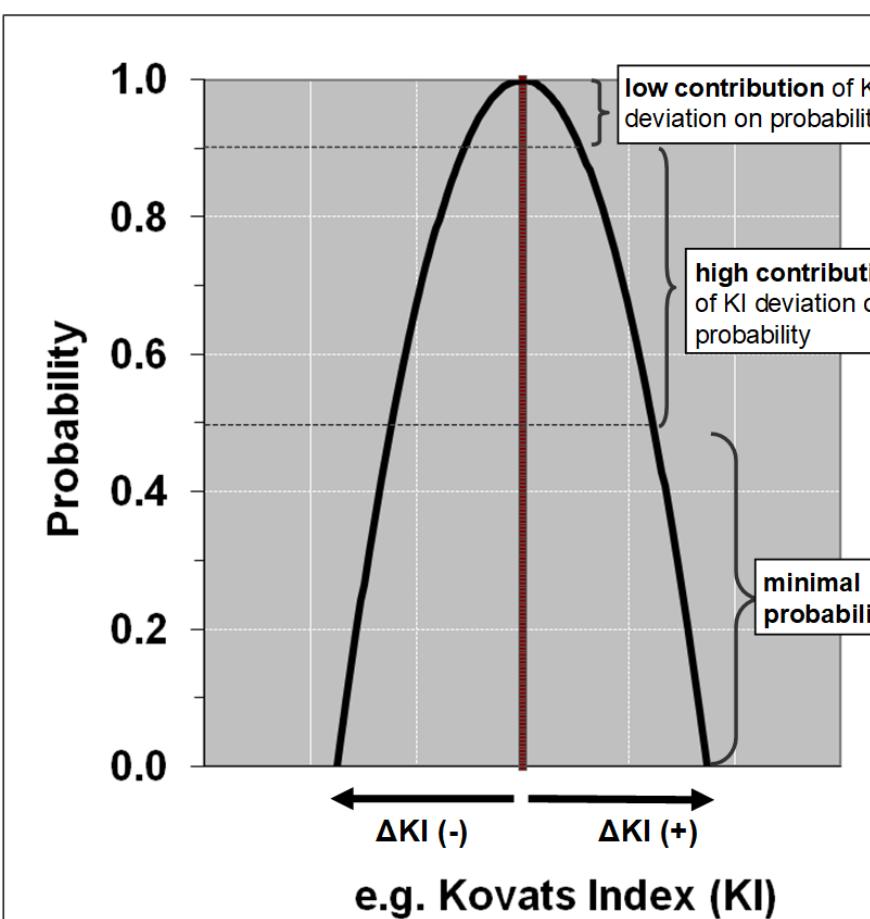
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

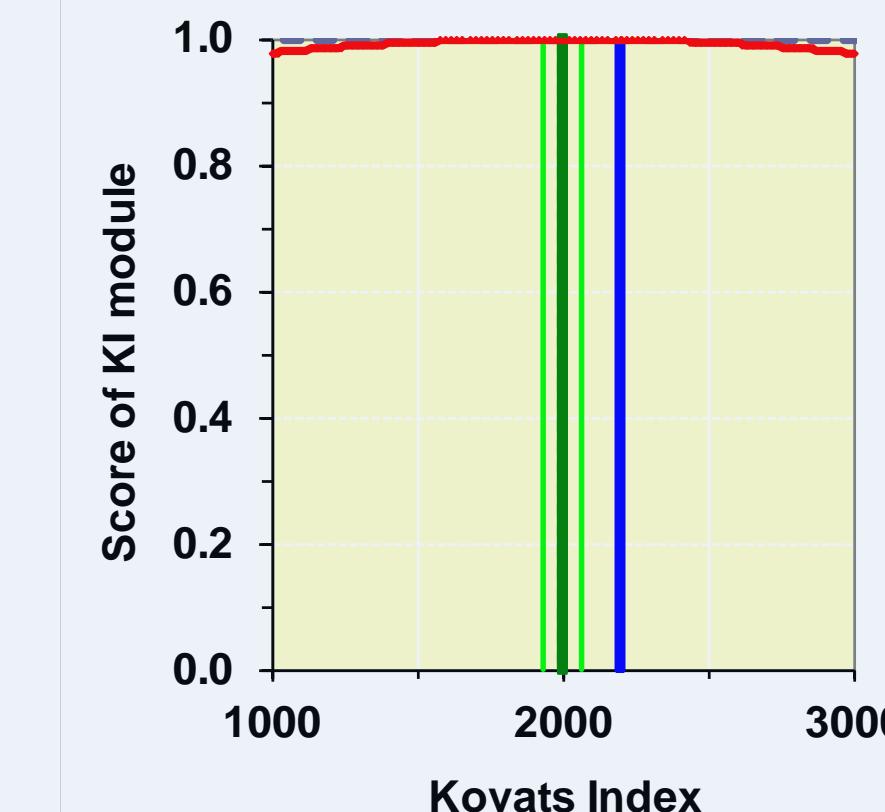
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	100.0
Score	0.999

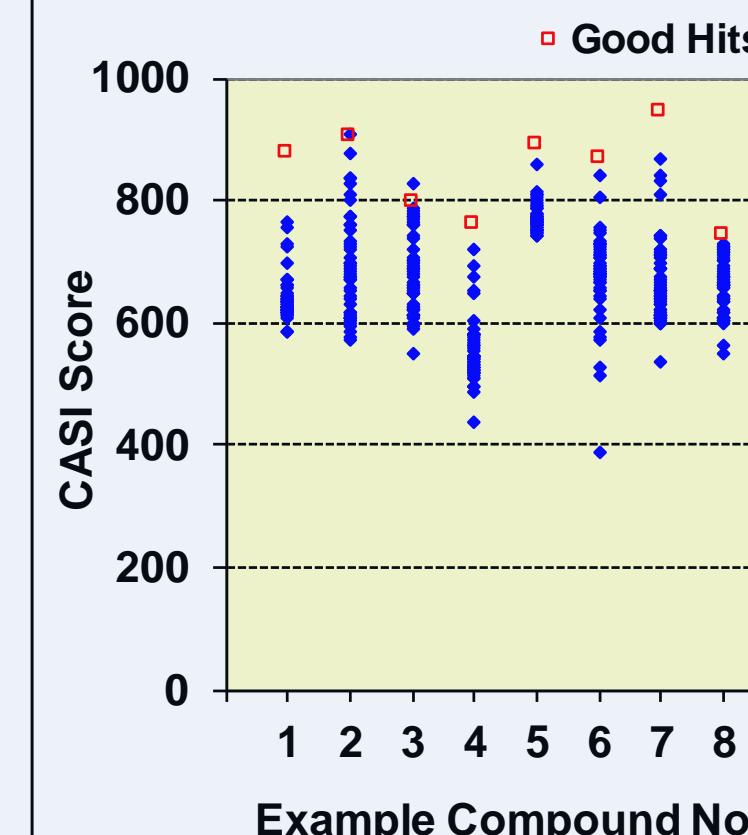
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

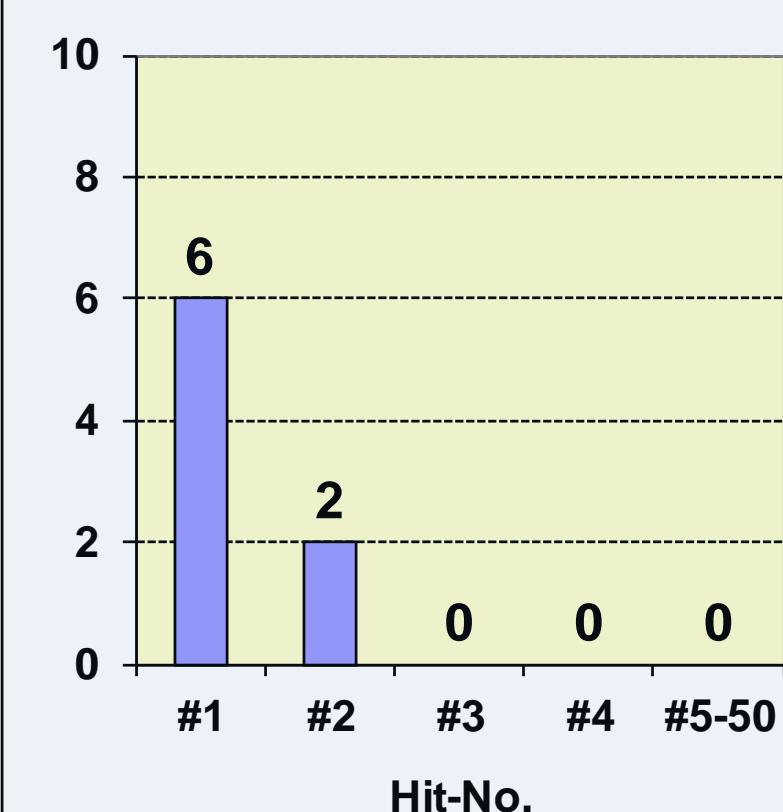
Visualization of curve fitting



Score by MS Similarity and Predicted KI



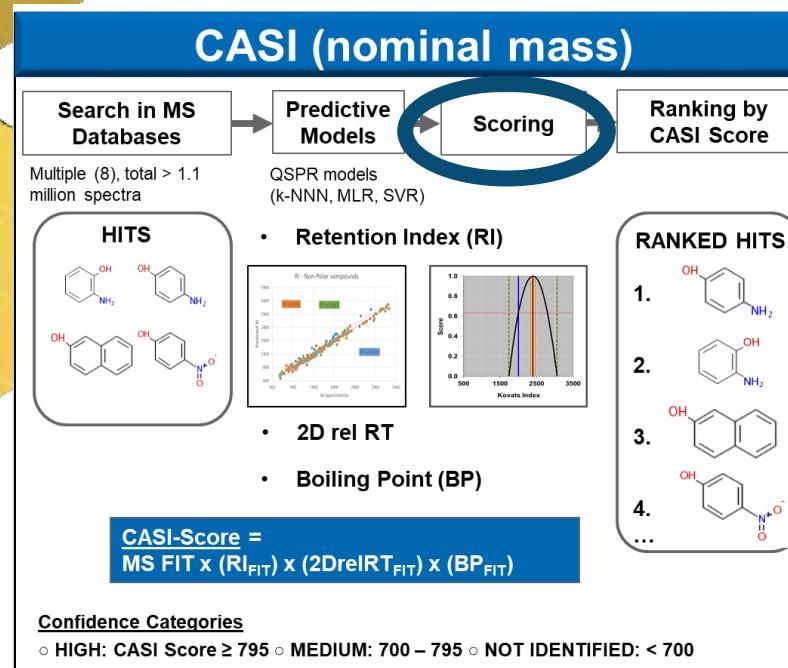
Hit ranking of correct structures



Score model optimization

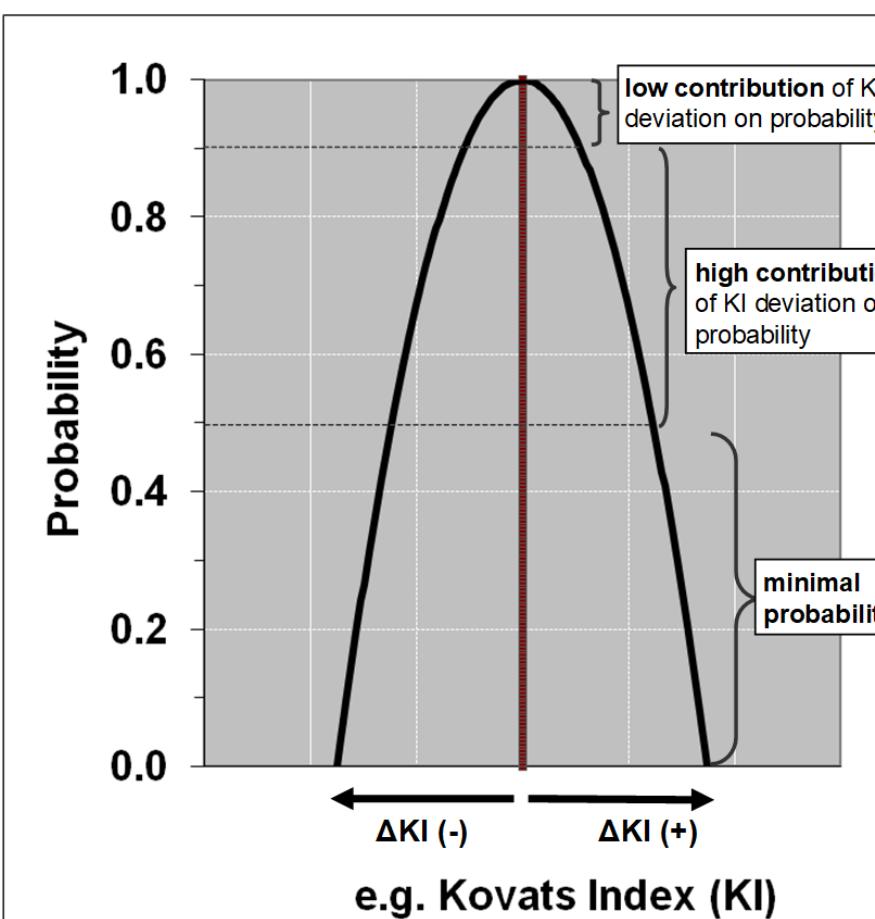
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

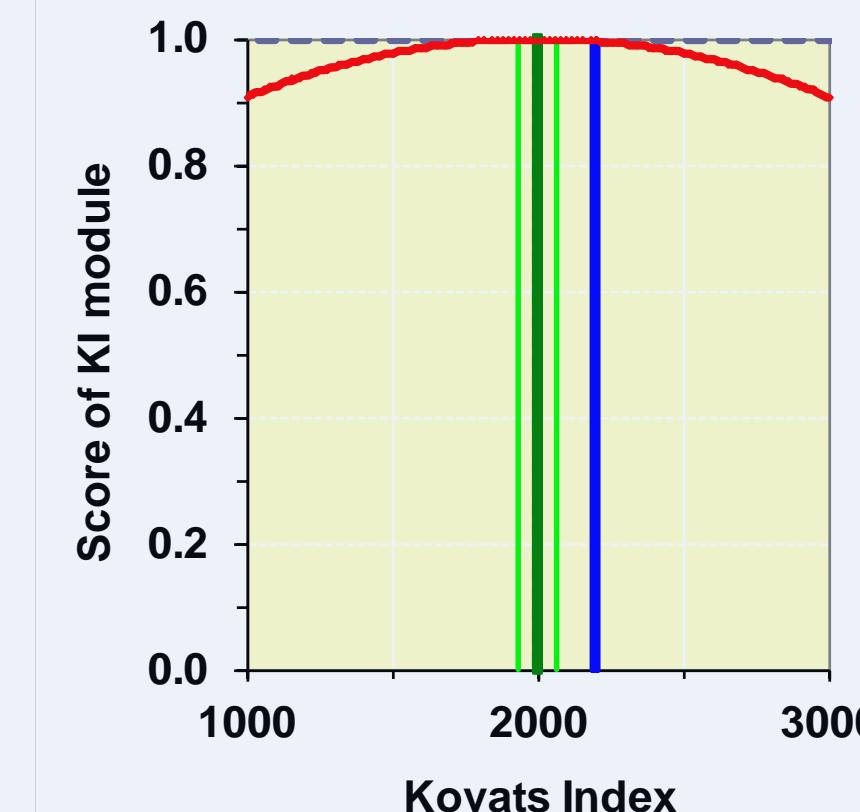
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	50.0
Score	0.996

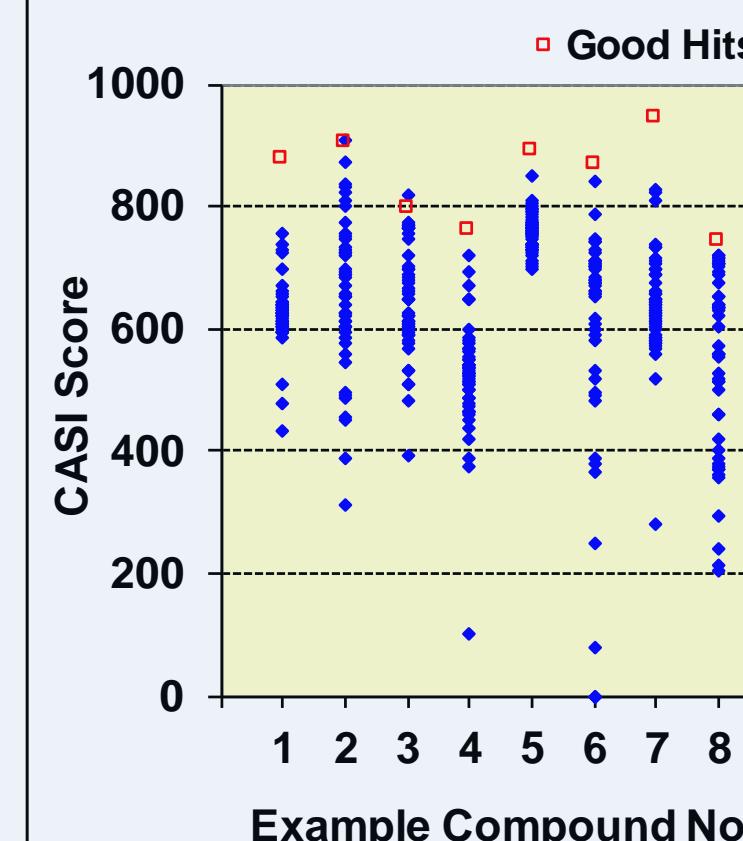
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

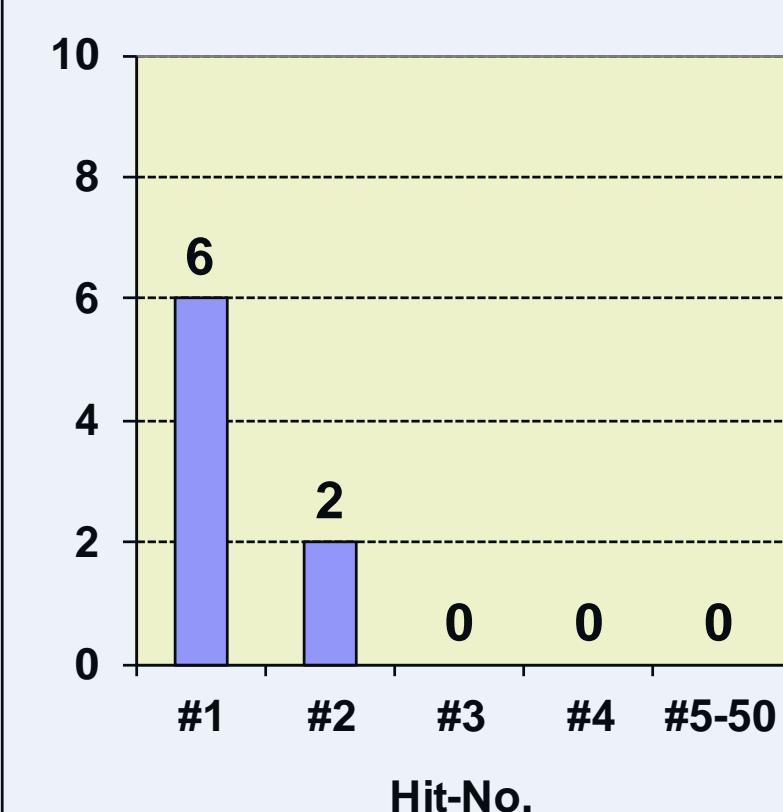
Visualization of curve fitting



Score by MS Similarity and Predicted KI



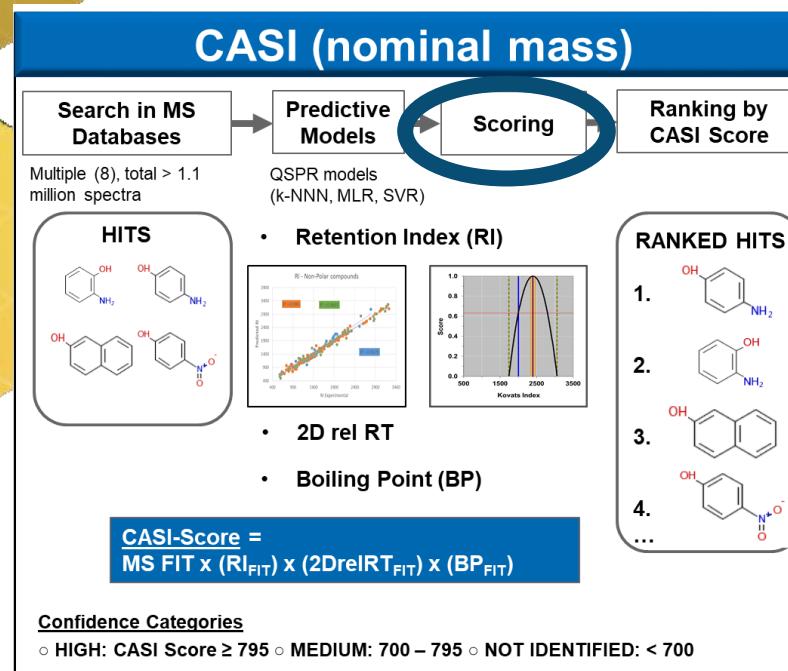
Hit ranking of correct structures



Score model optimization

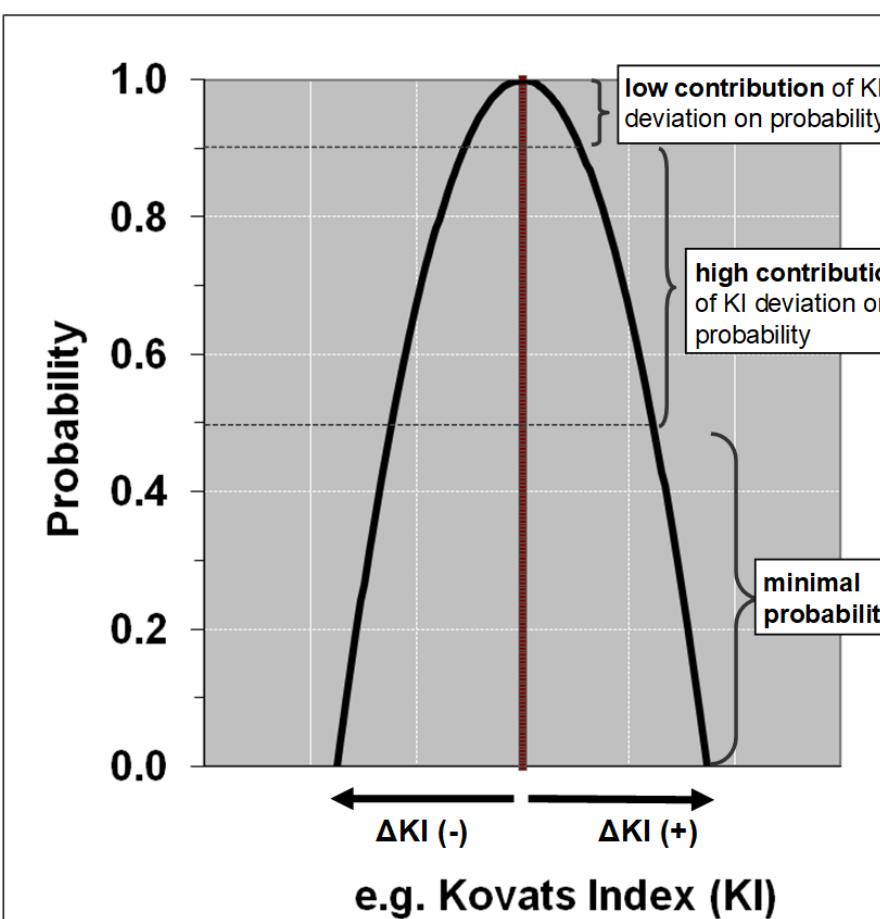
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

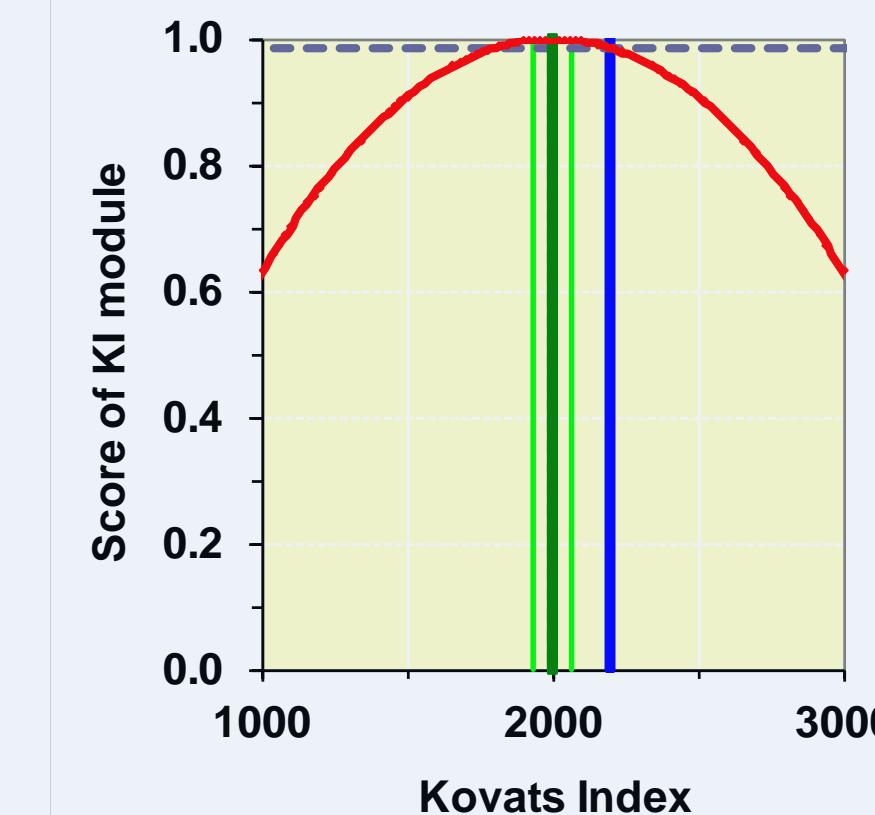
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	25.0
Score	0.985

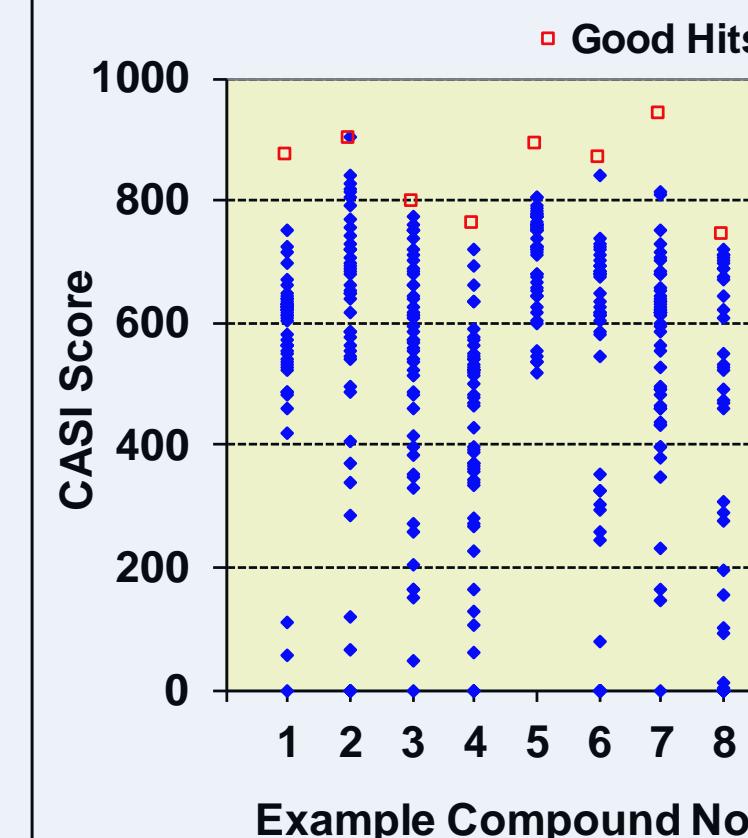
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

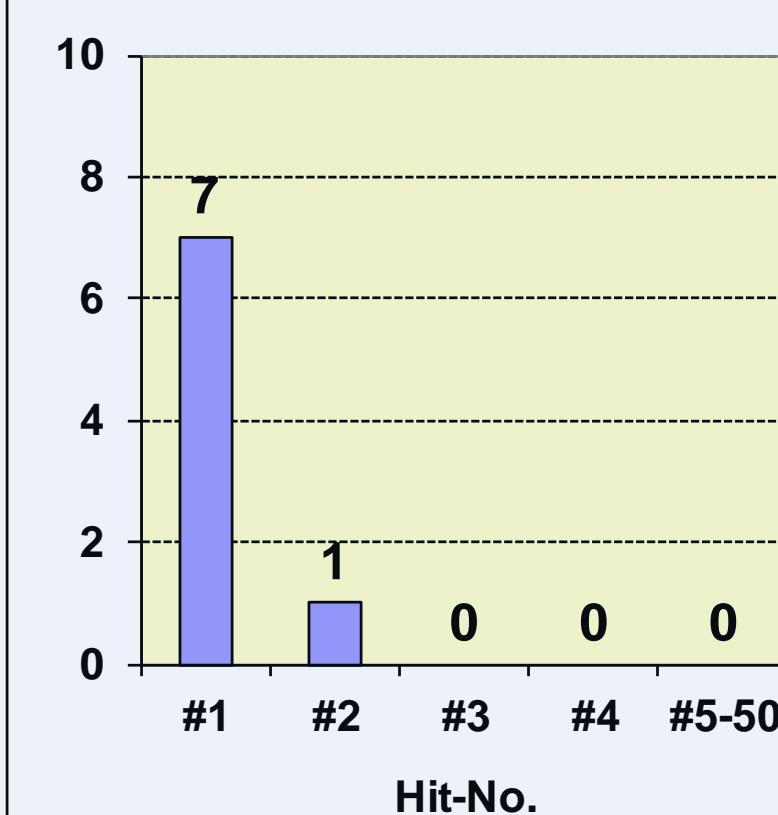
Visualization of curve fitting



Score by MS Similarity and Predicted KI



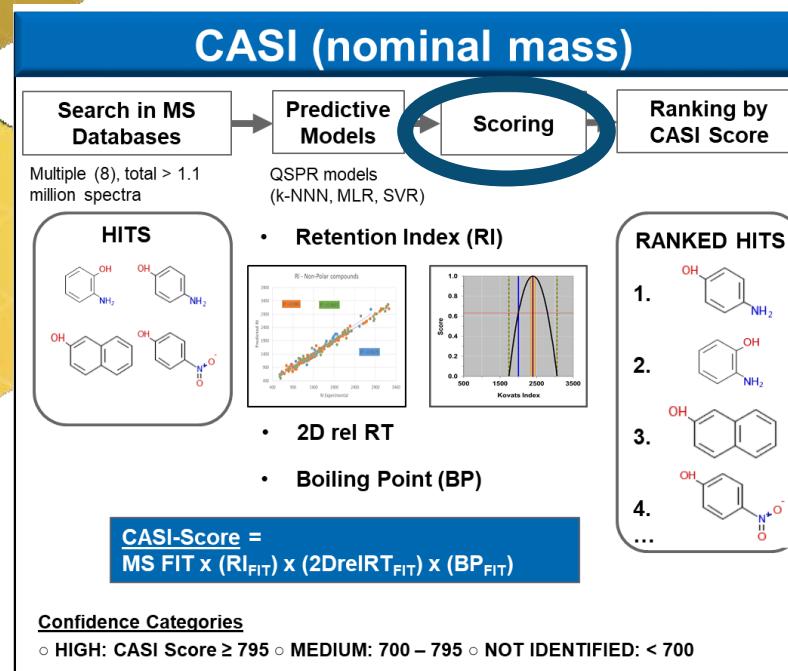
Hit ranking of correct structures



Score model optimization

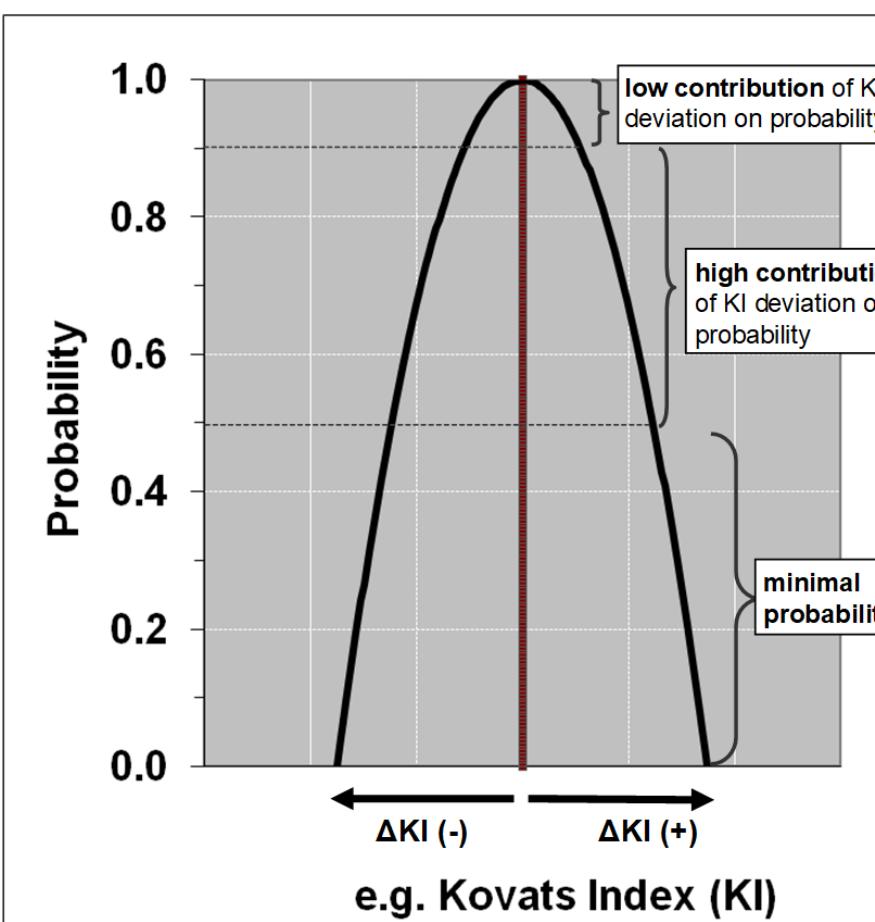
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

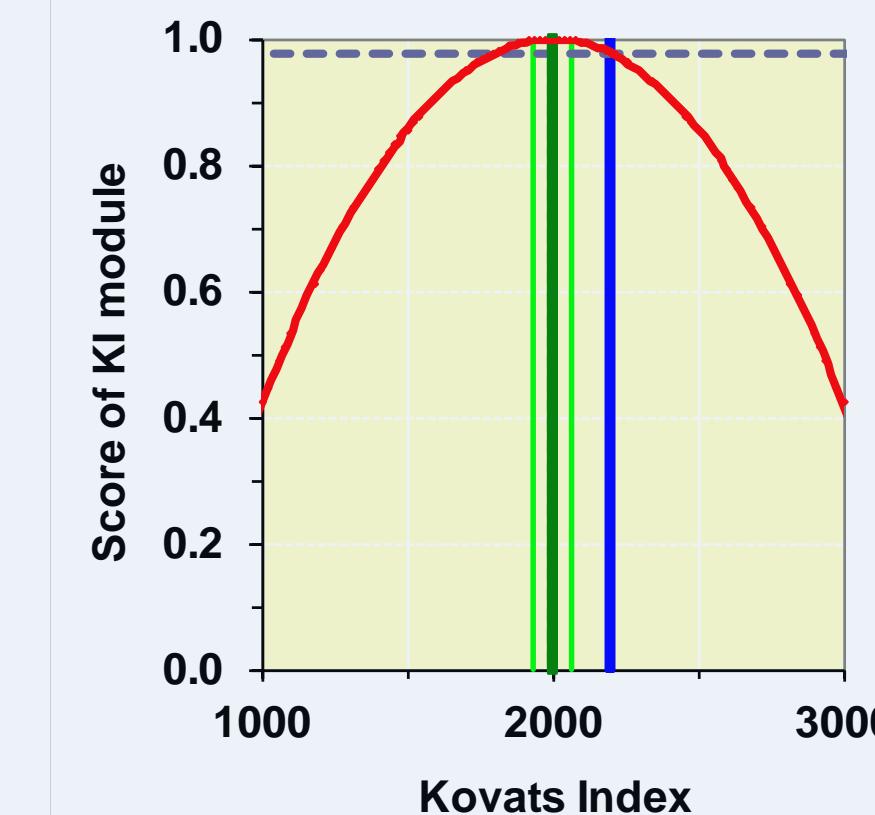
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	20.0
Score	0.977

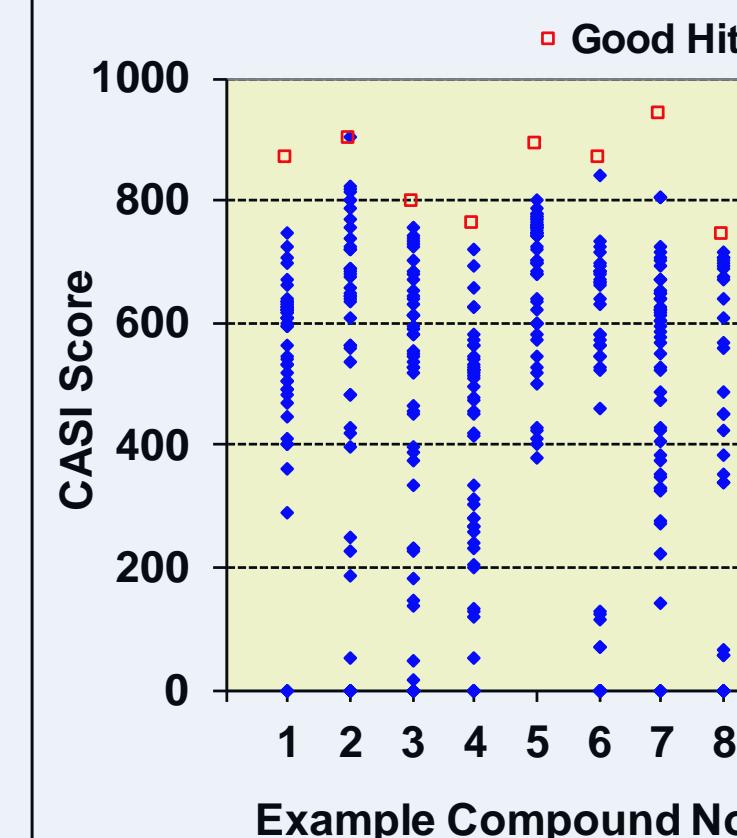
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

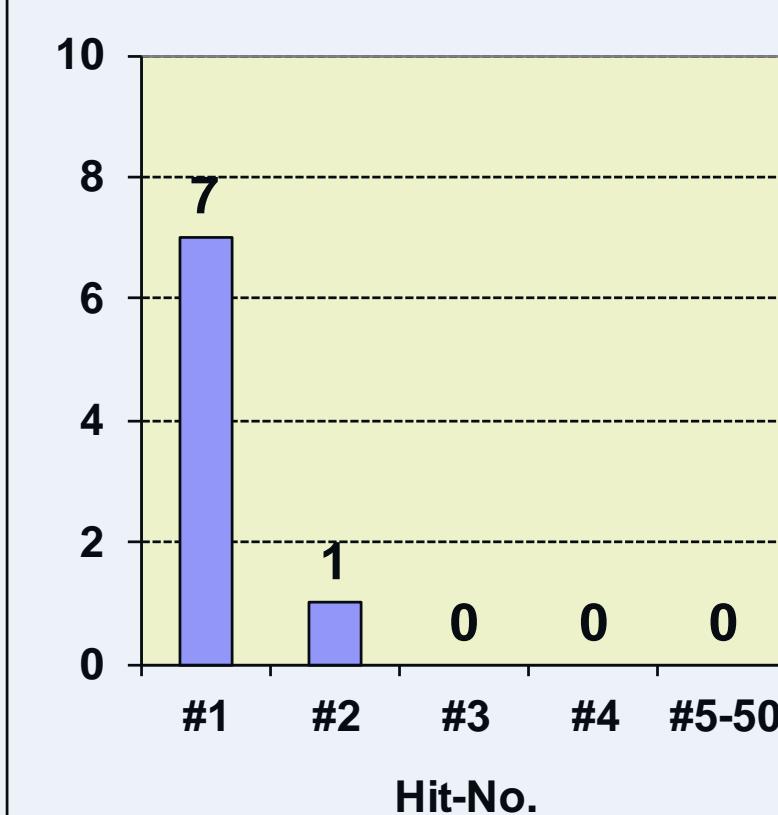
Visualization of curve fitting



Score by MS Similarity and Predicted KI



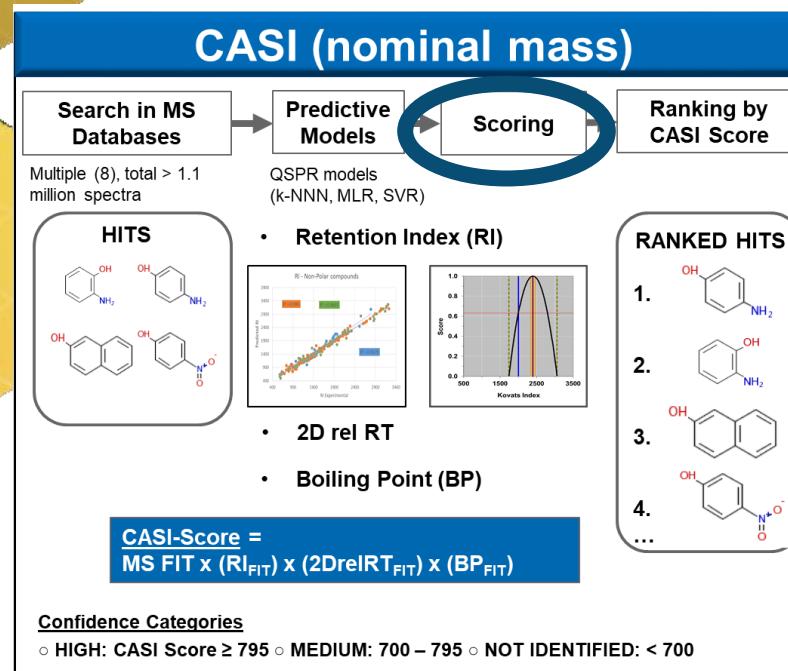
Hit ranking of correct structures



Score model optimization

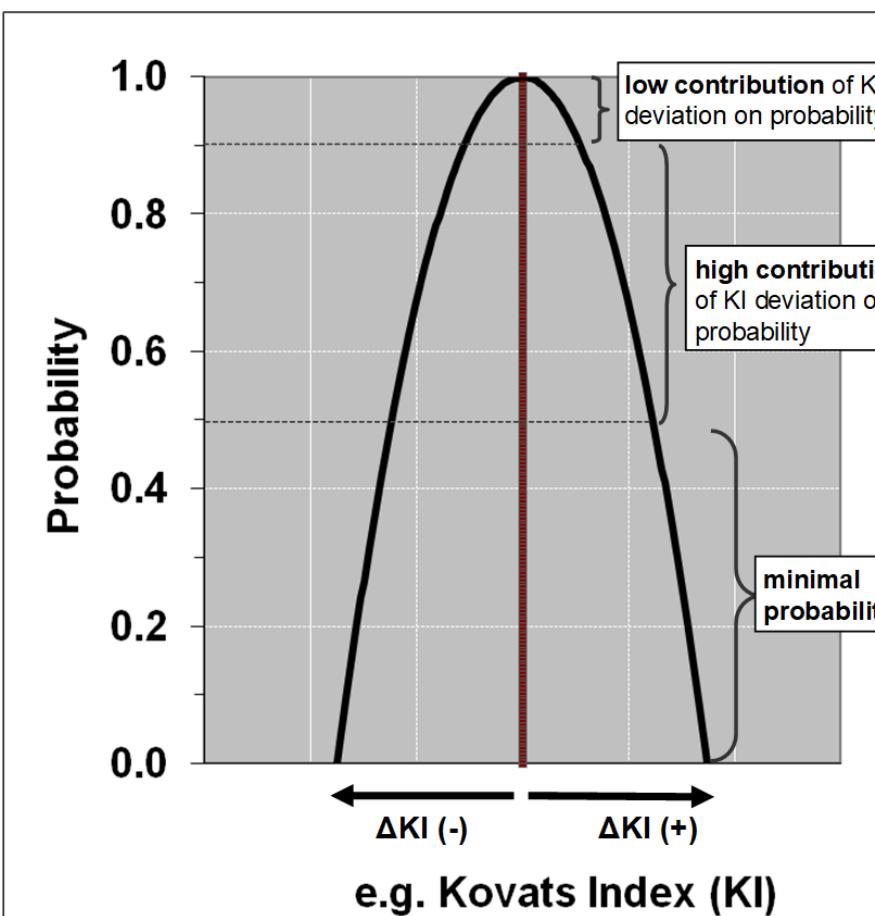
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

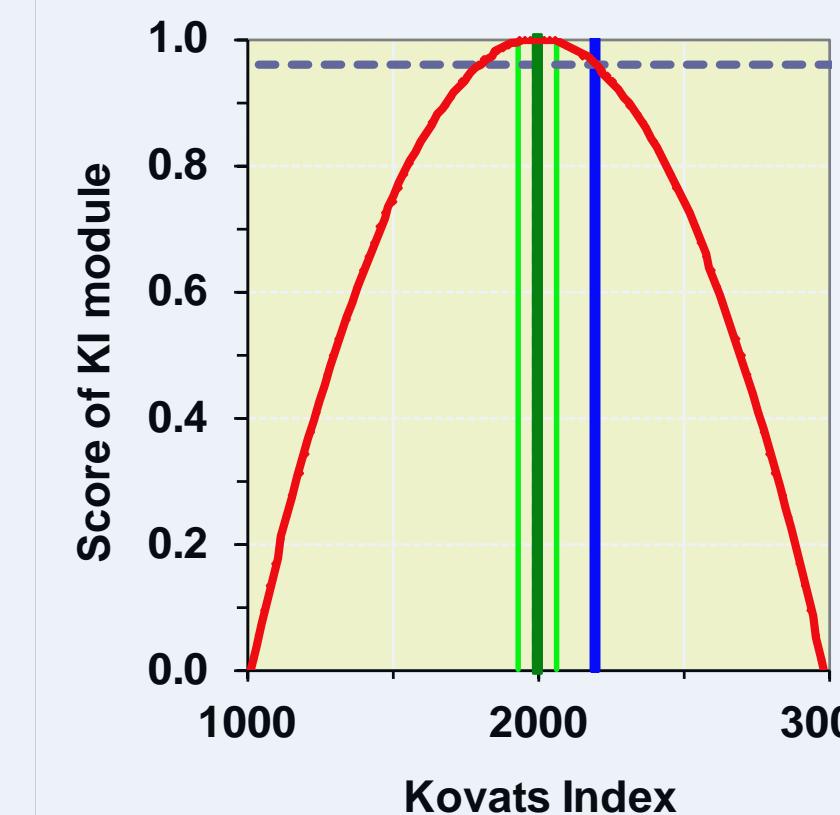
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	15.0
Score	0.959

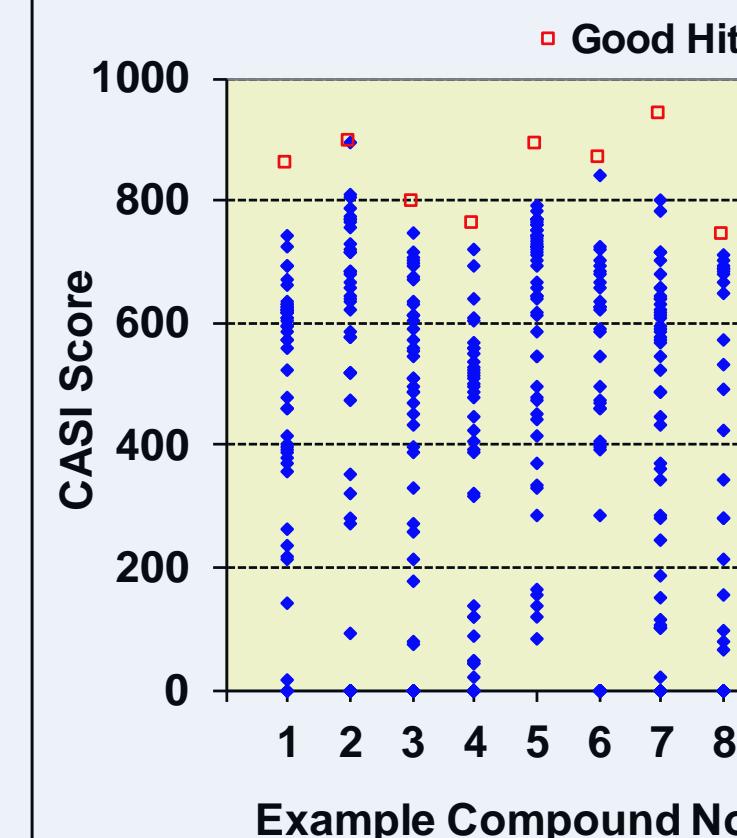
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

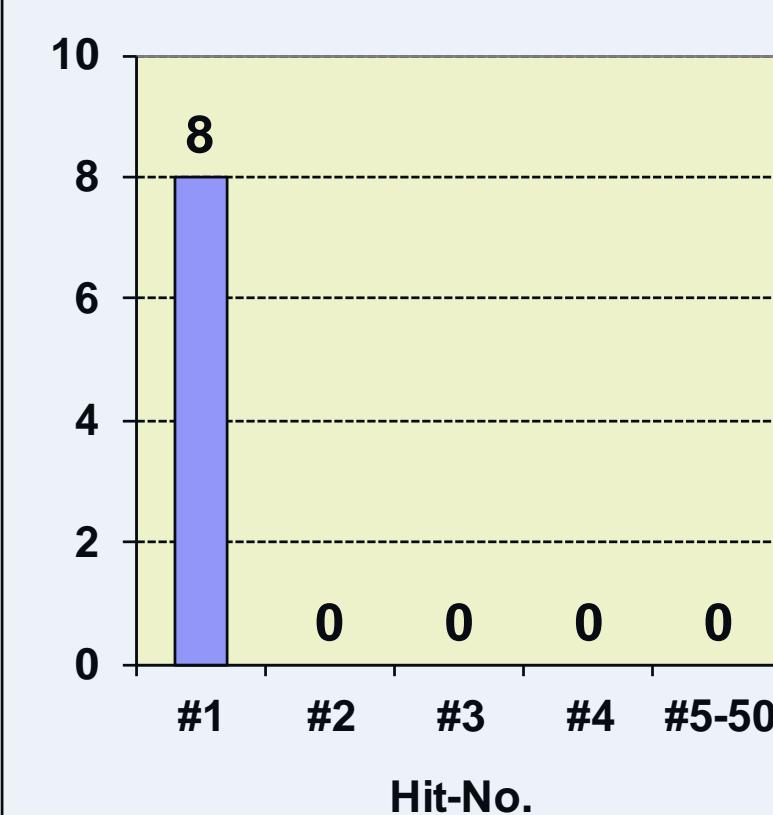
Visualization of curve fitting



Score by MS Similarity and Predicted KI



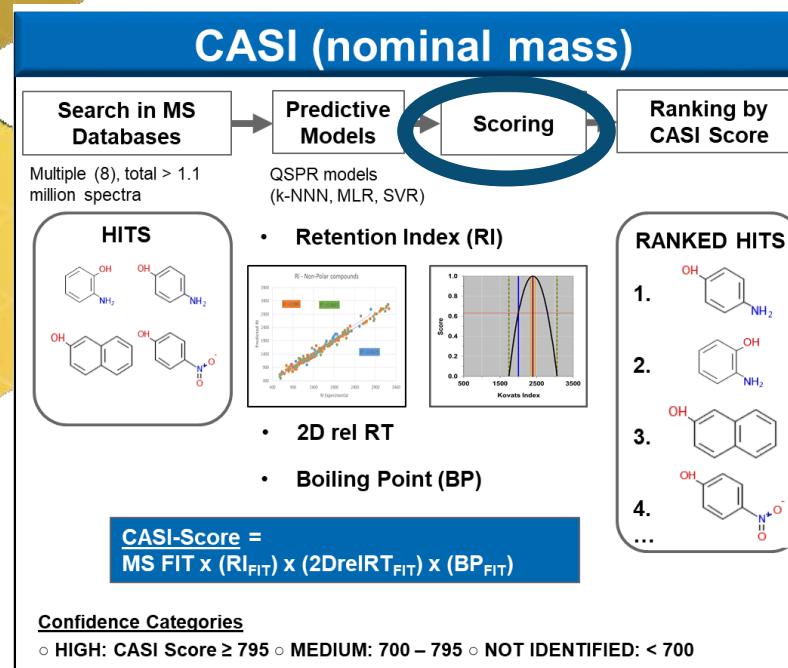
Hit ranking of correct structures



Score model optimization

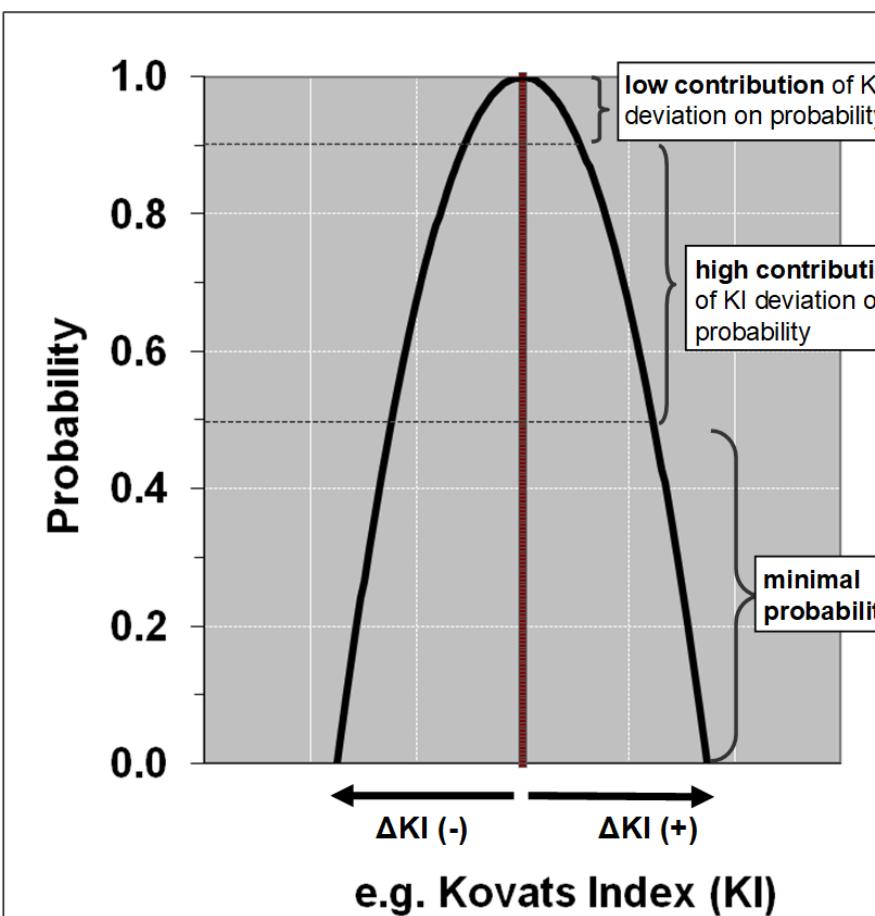
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

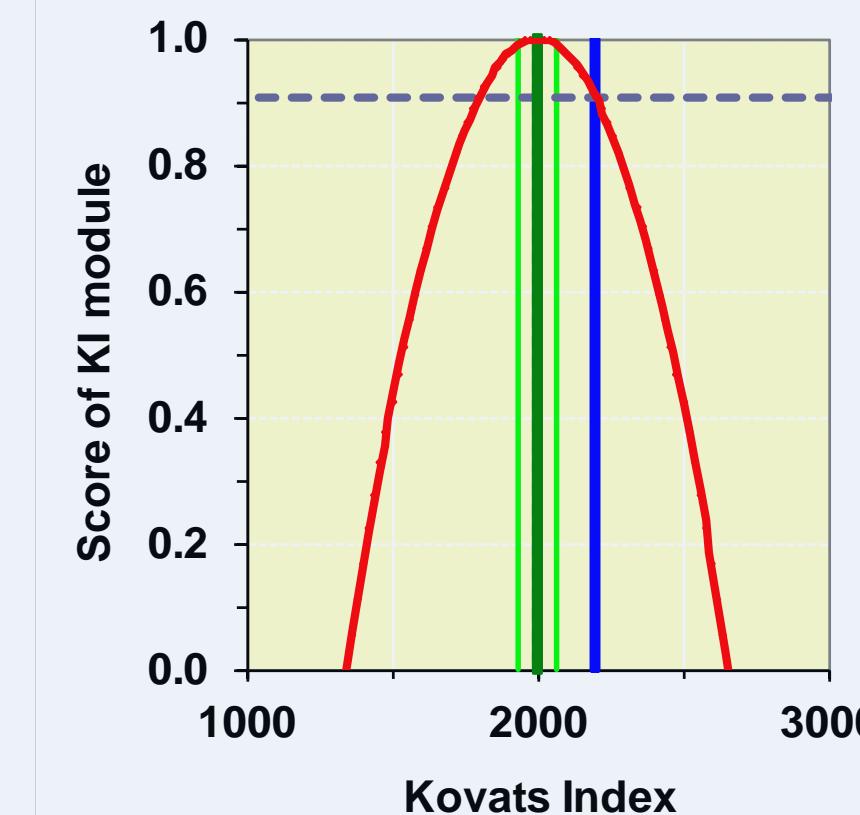
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	10.0
Score	0.908

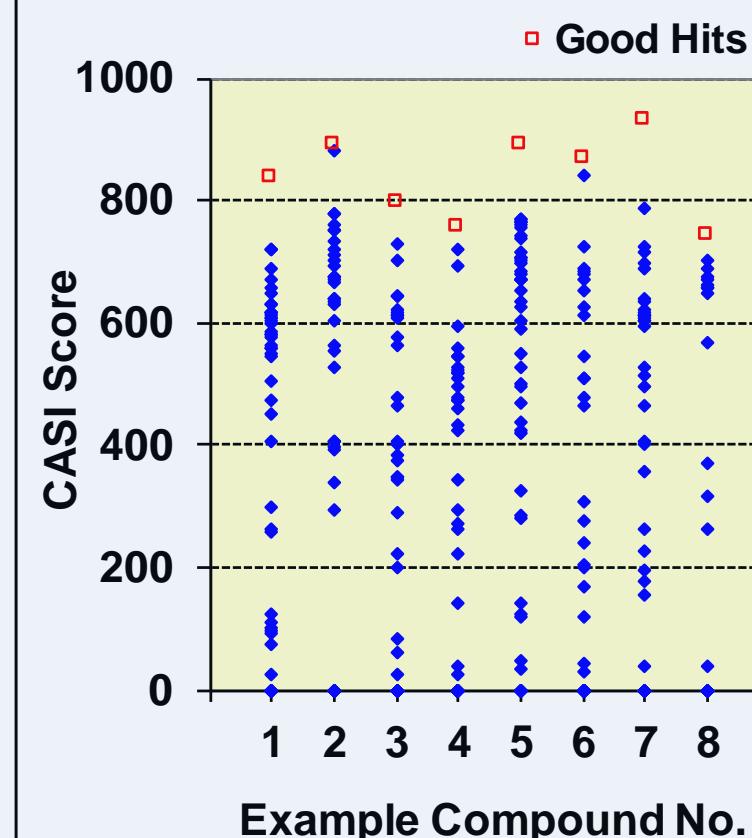
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

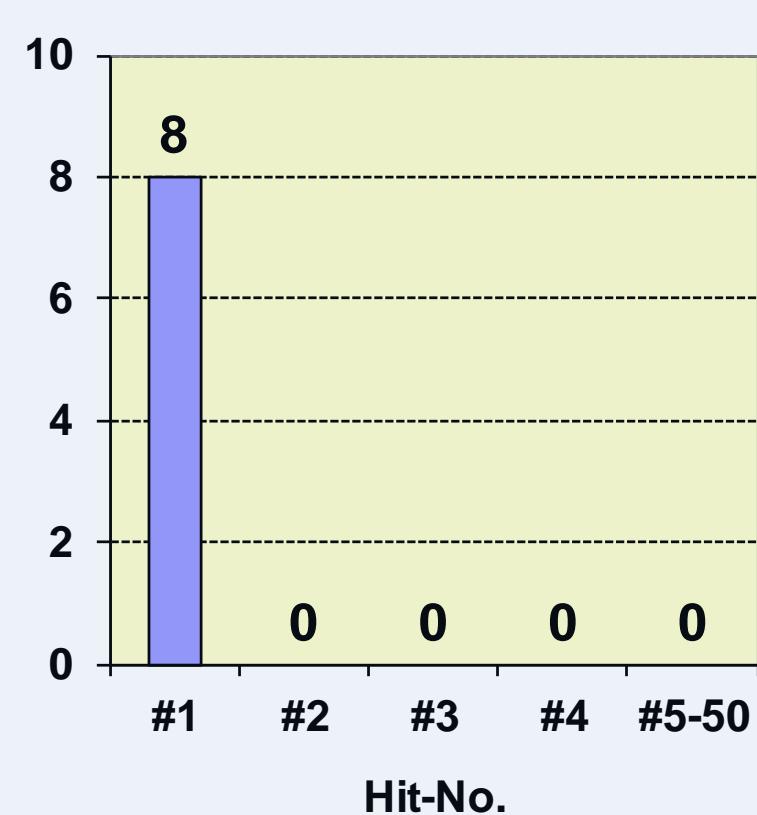
Visualization of curve fitting



Score by MS Similarity and Predicted KI



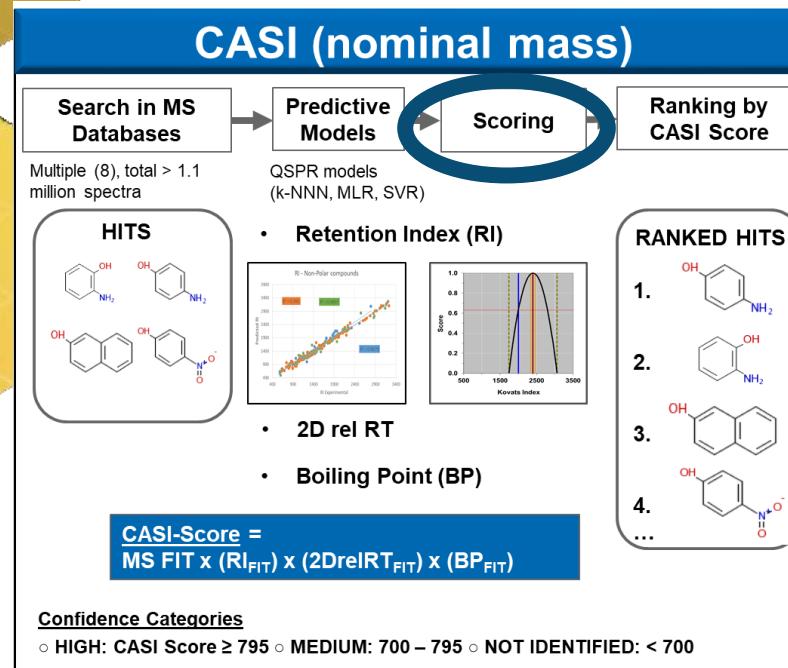
Hit ranking of correct structures



Score model optimization

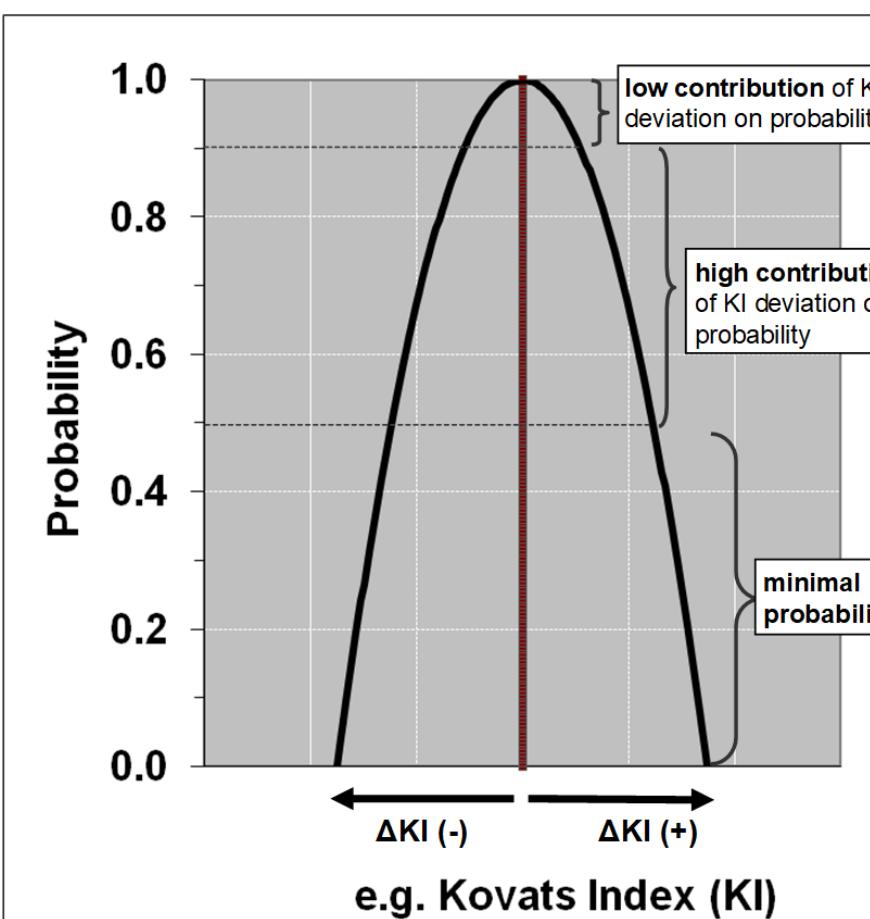
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

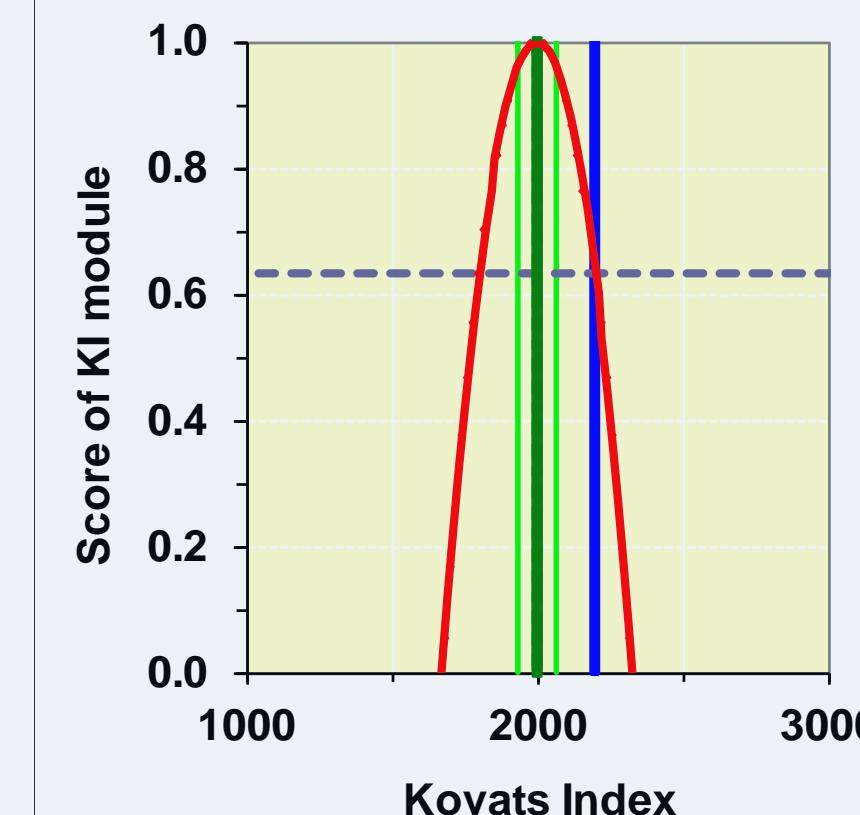
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	5.0
Score	0.630

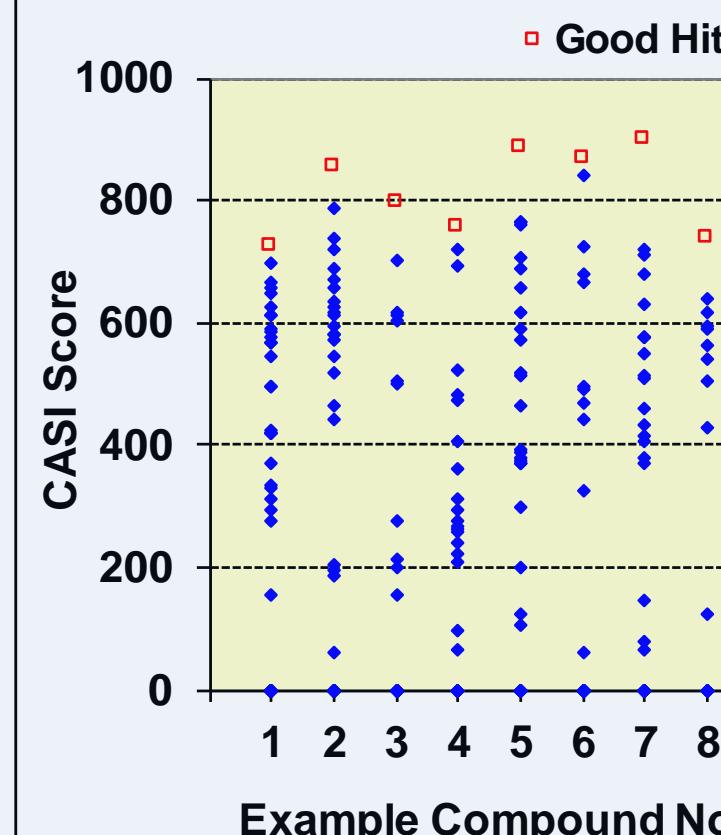
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

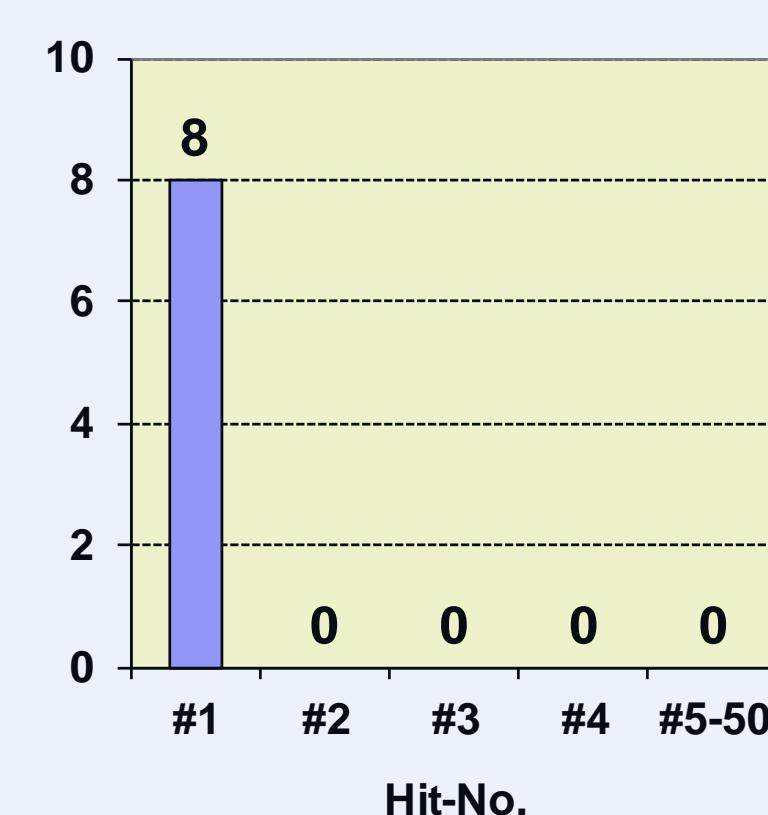
Visualization of curve fitting



Score by MS Similarity and Predicted KI



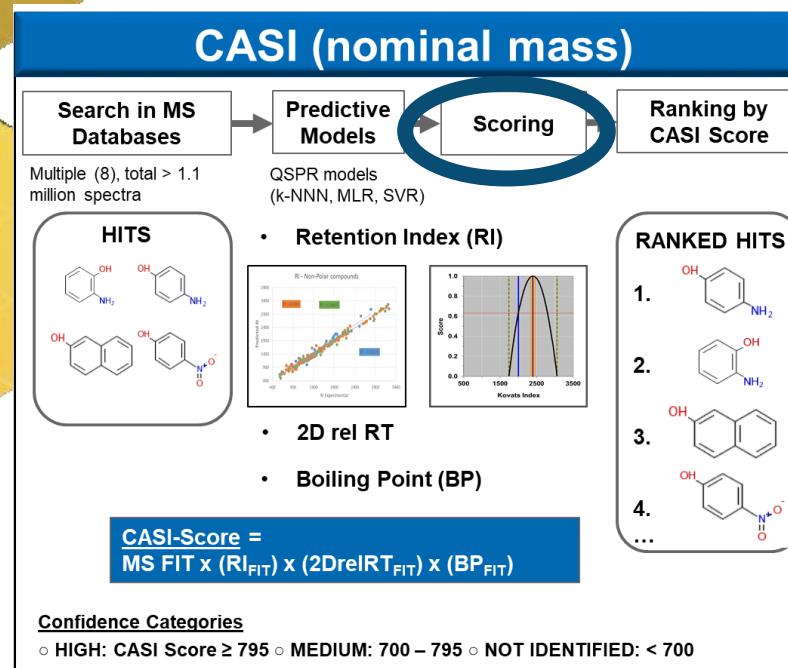
Hit ranking of correct structures



Score model optimization

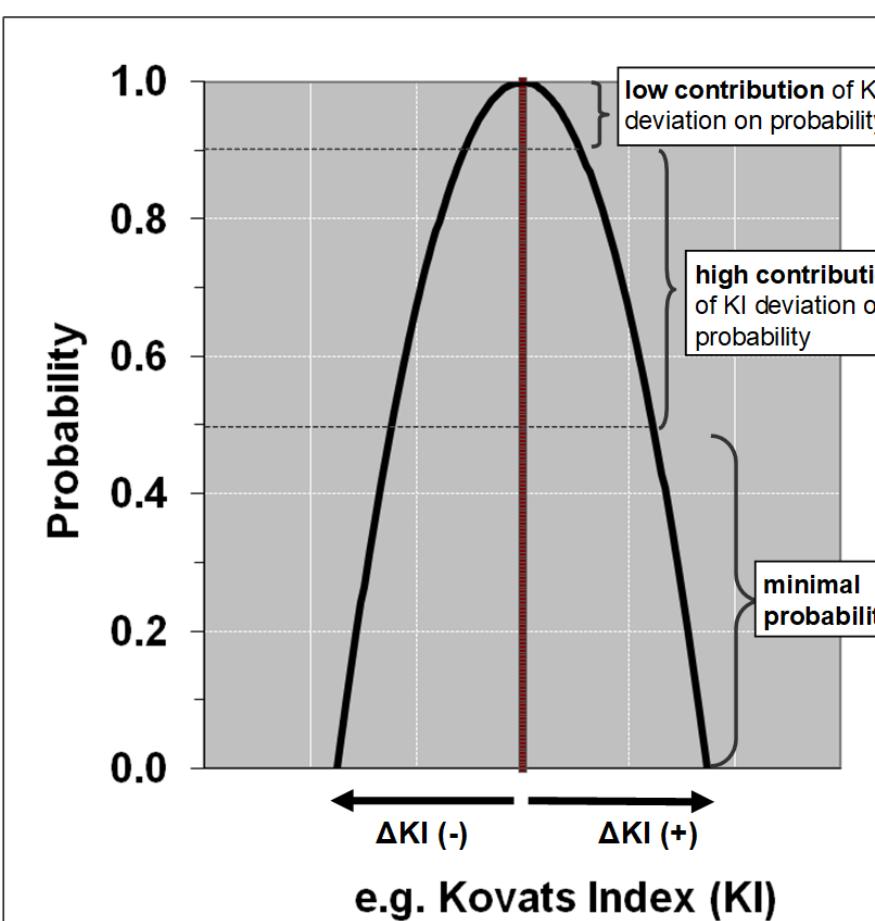
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

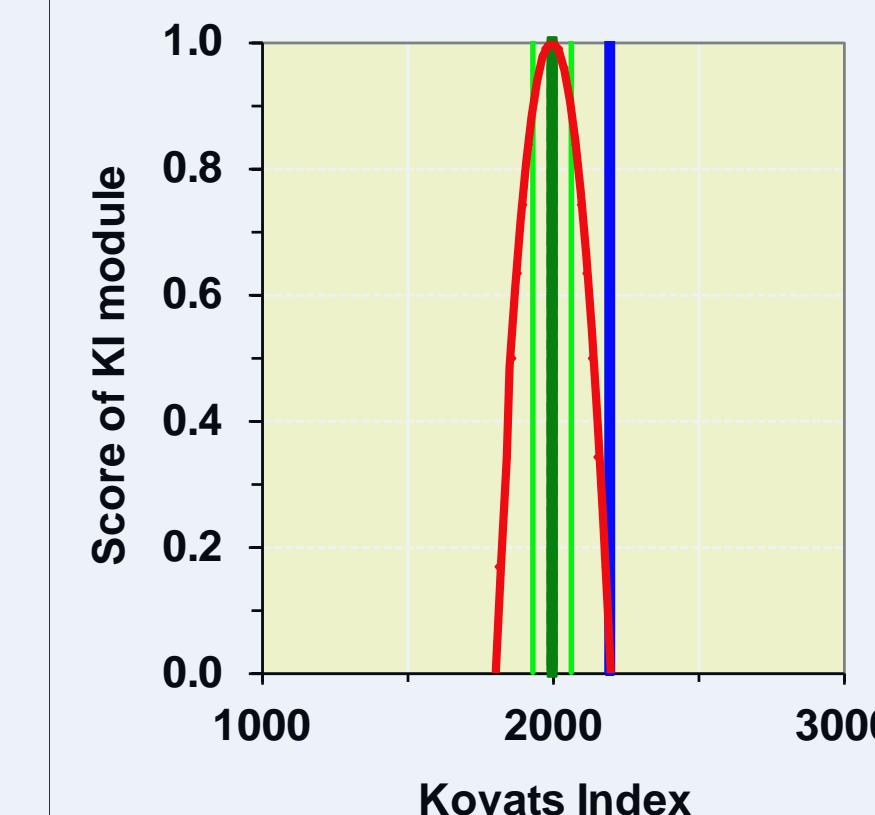
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	3.0
Score	0.000

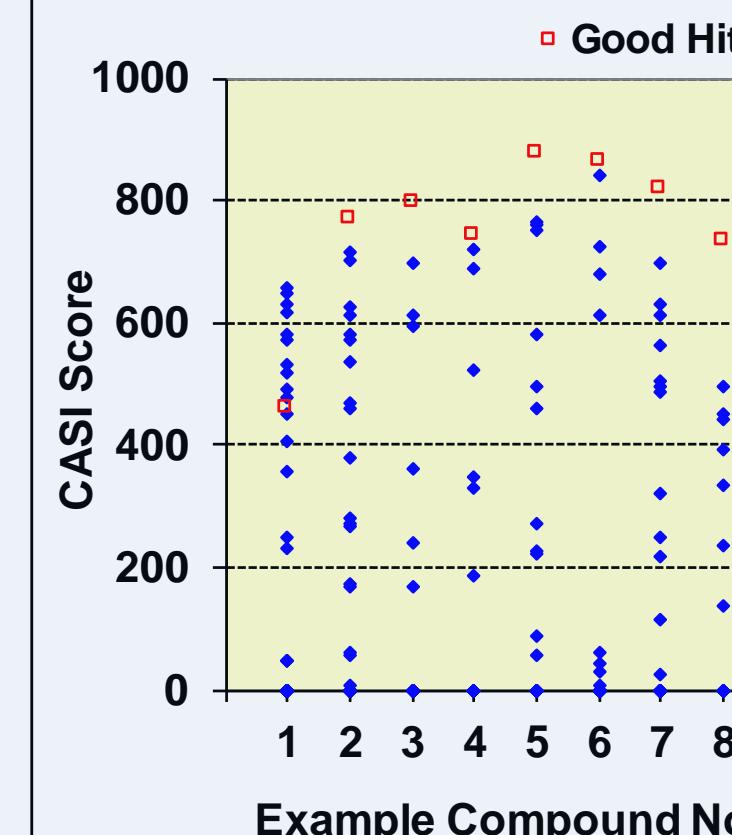
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

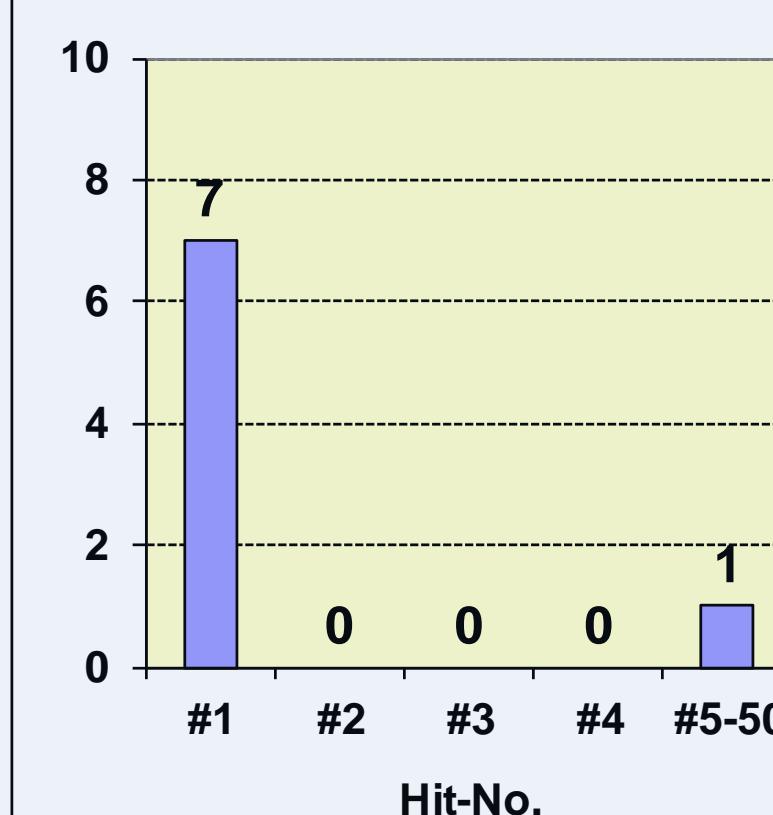
Visualization of curve fitting



Score by MS Similarity and Predicted KI



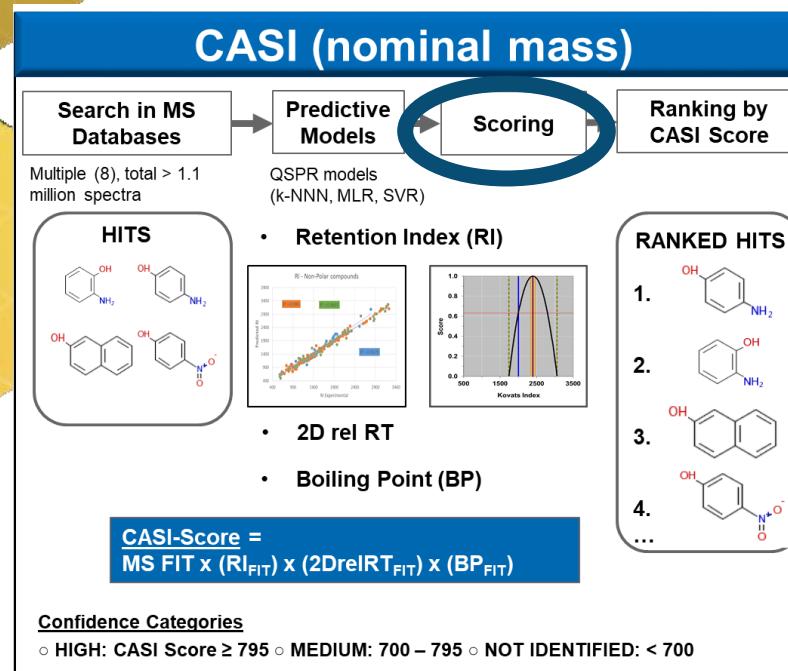
Hit ranking of correct structures



Score model optimization

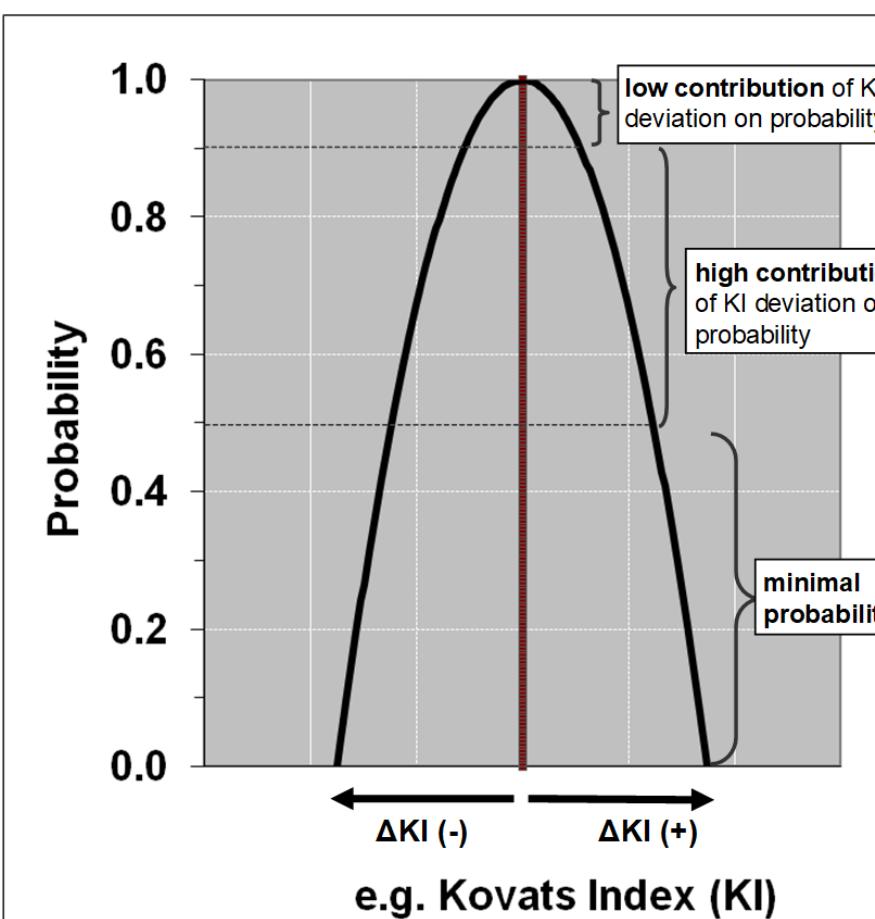
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

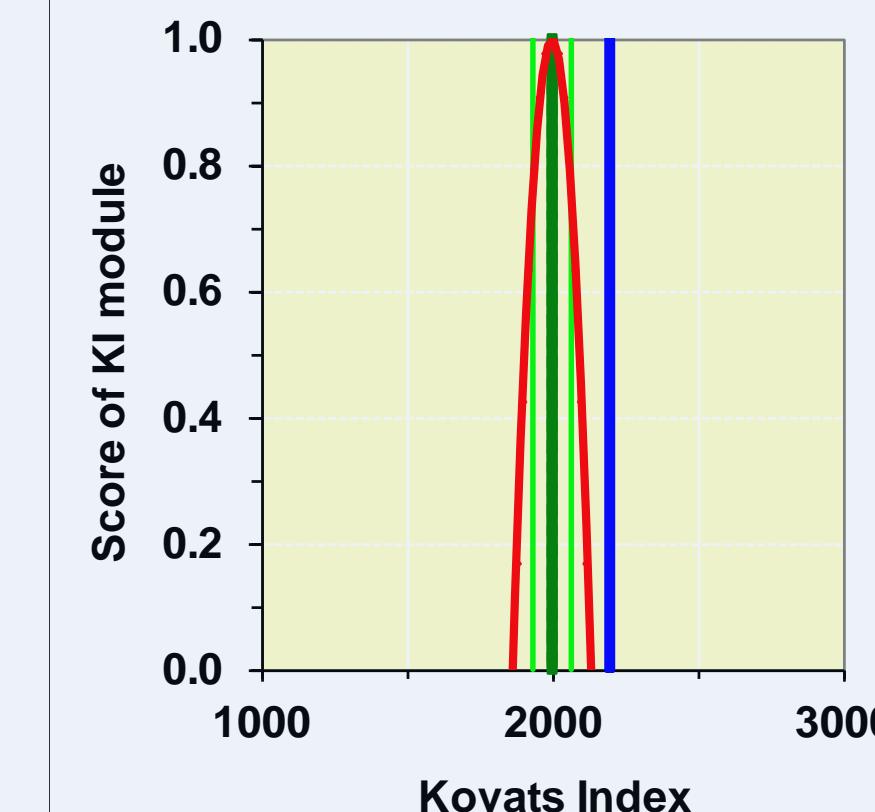
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	2.0
Score	0.000

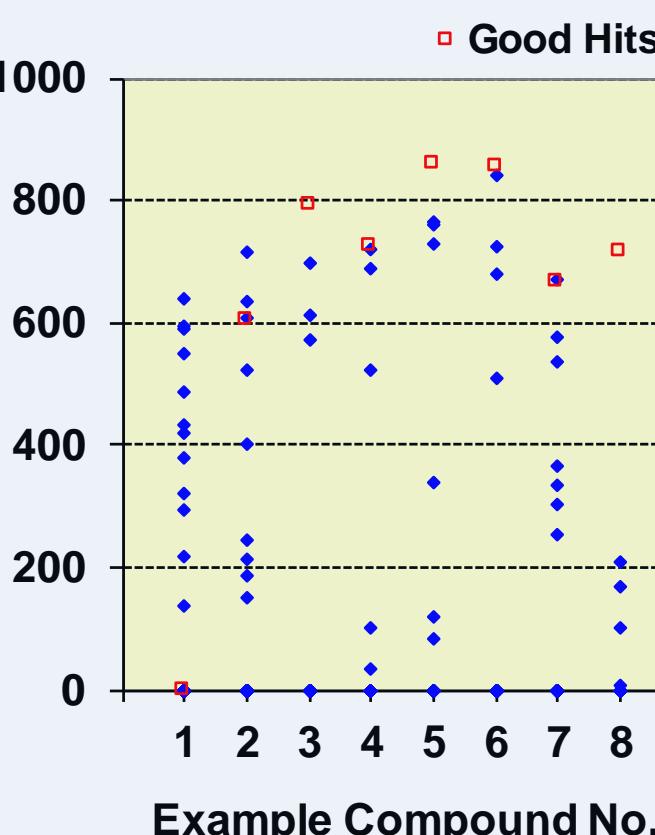
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

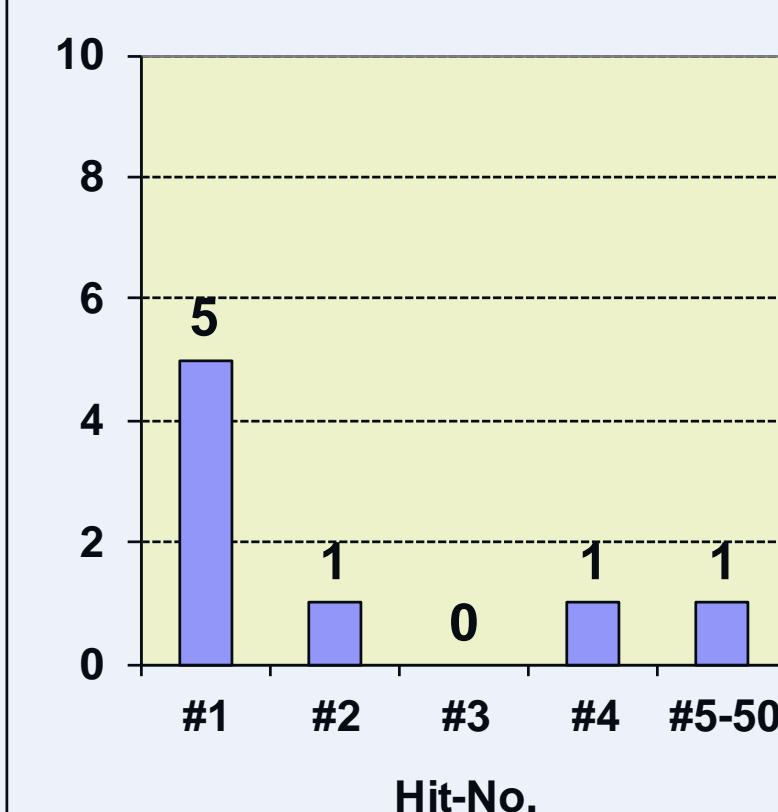
Visualization of curve fitting



Score by MS Similarity and Predicted KI



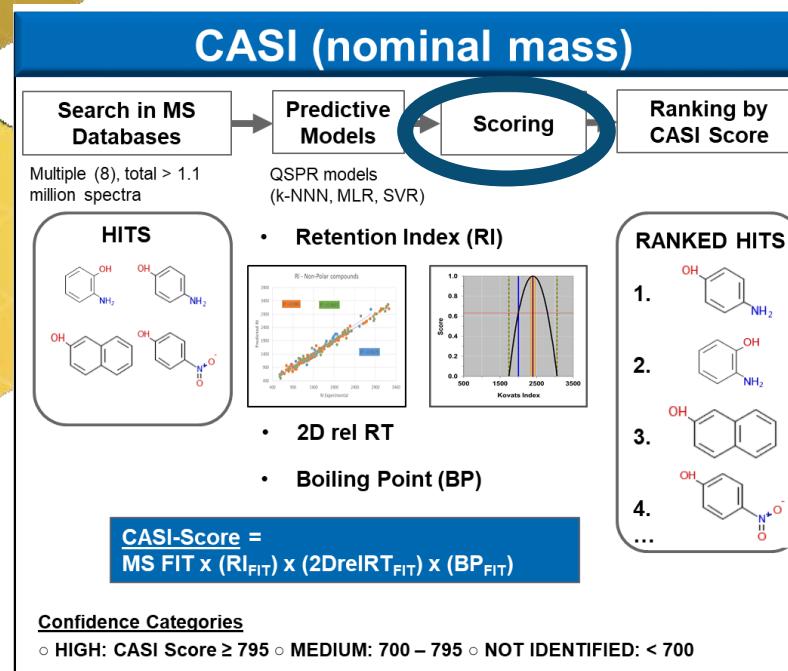
Hit ranking of correct structures



Score model optimization

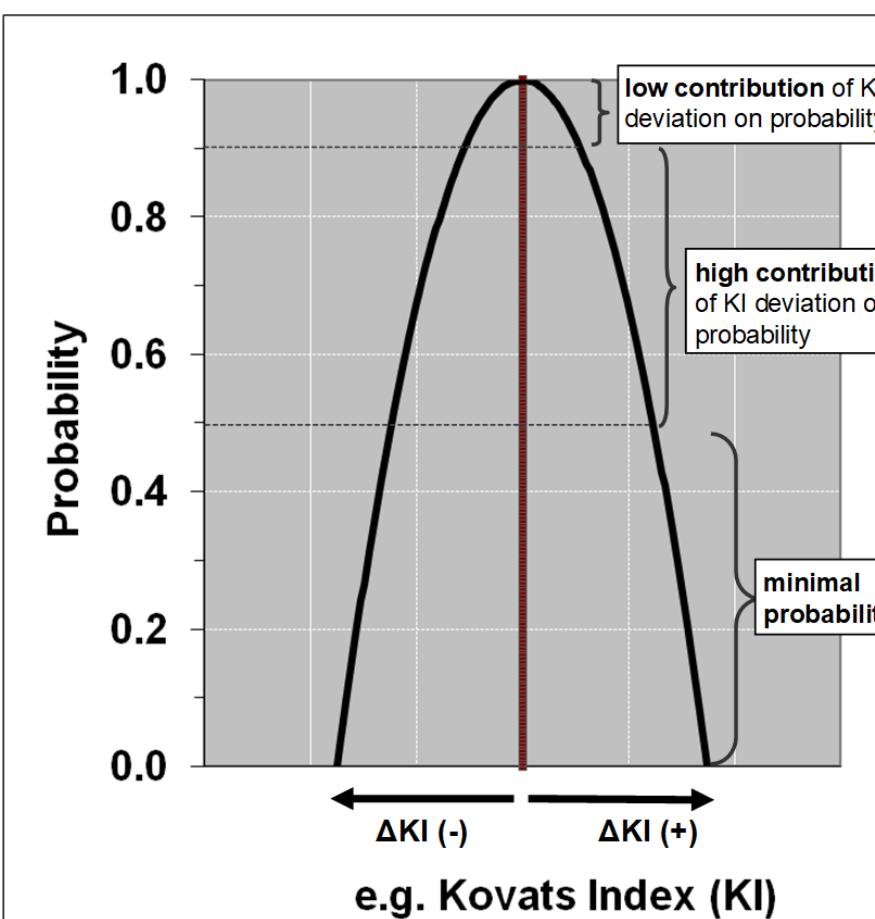
- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

High-throughput structure identification (GC — Nominal mass)



Score functions

- Parabolic function for each analytical property



Curve Visualization

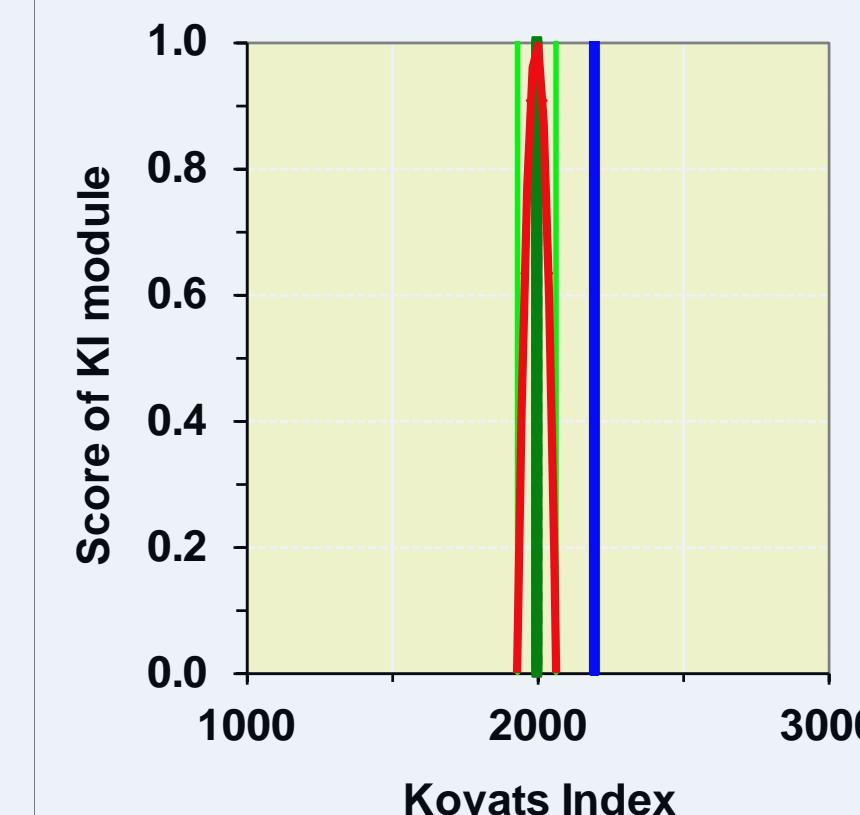
Example: 8x50 structures with highest NIST FIT

Predicted KI	2000.0
Experimental KI	2200.0
SE Prediction	65.8
Curve fitting (variable)	1.0
Score	0.000

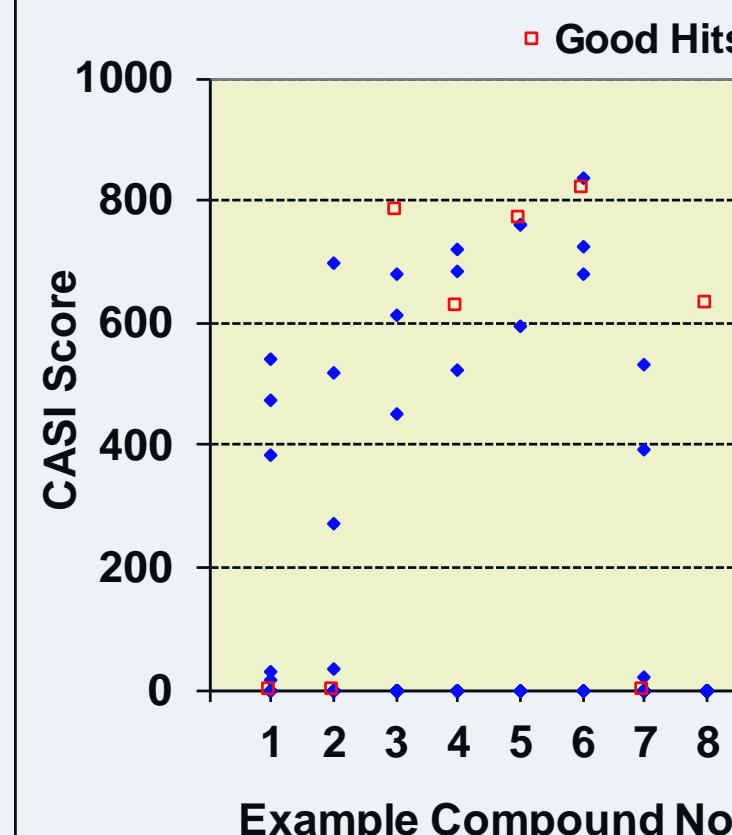
Red squares: confirmed structure

Blue dots: Incorrect proposals from NIST

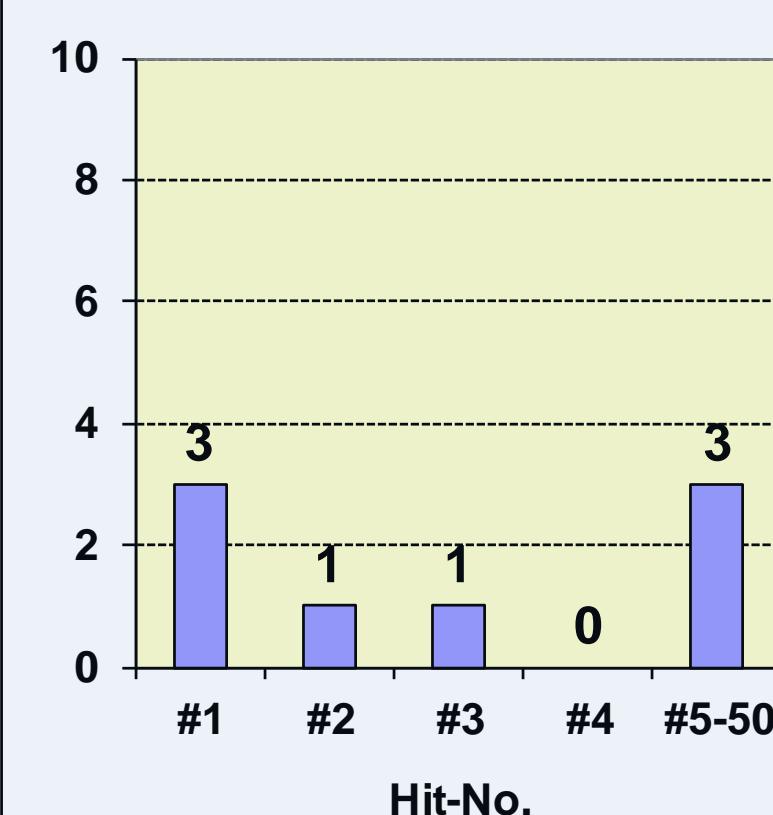
Visualization of curve fitting



Score by MS Similarity and Predicted KI



Hit ranking of correct structures



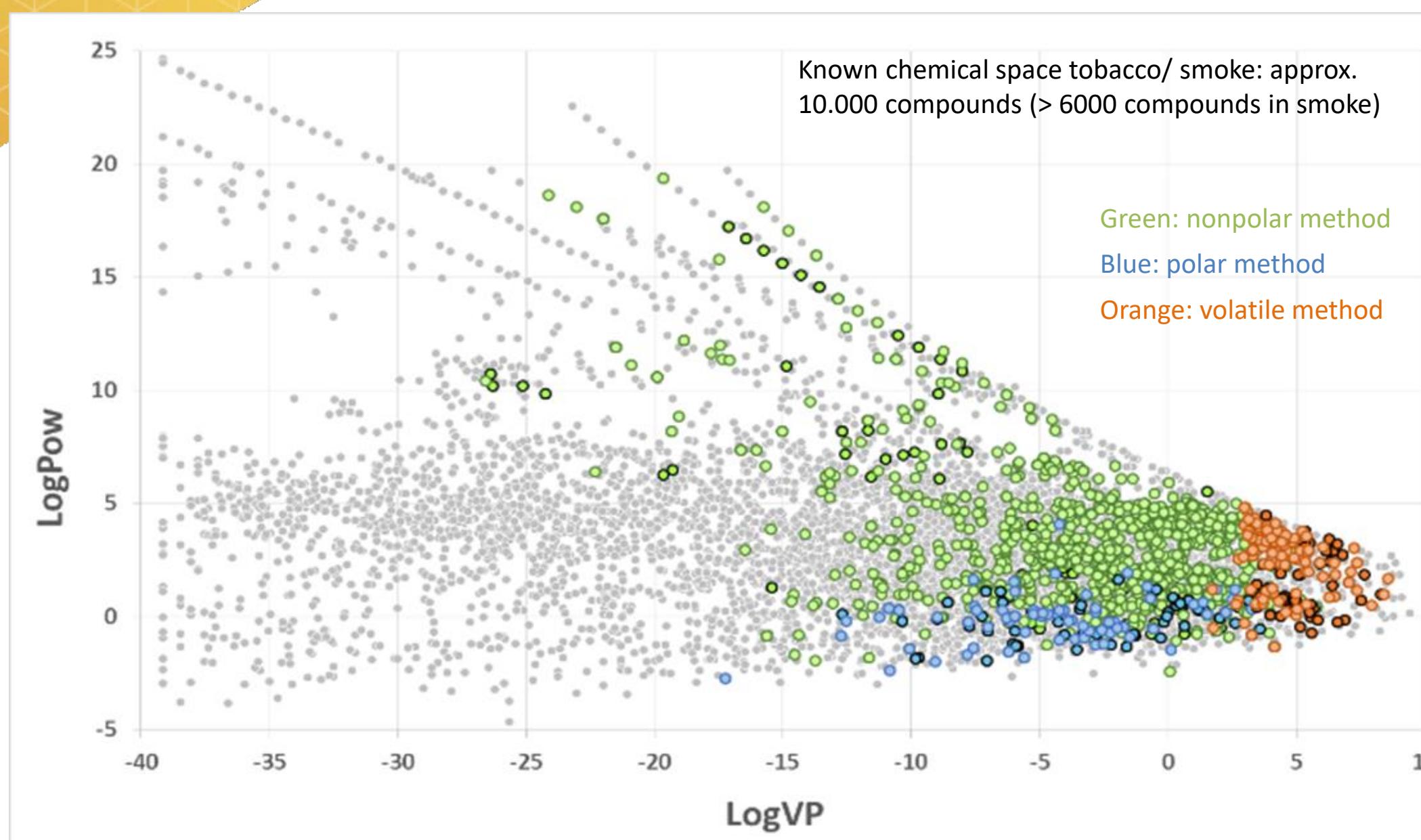
Score model optimization

- All equations on the fly (grid-matrix analysis) for result optimization (learning set)

Performance CASI (GC — nominal mass)

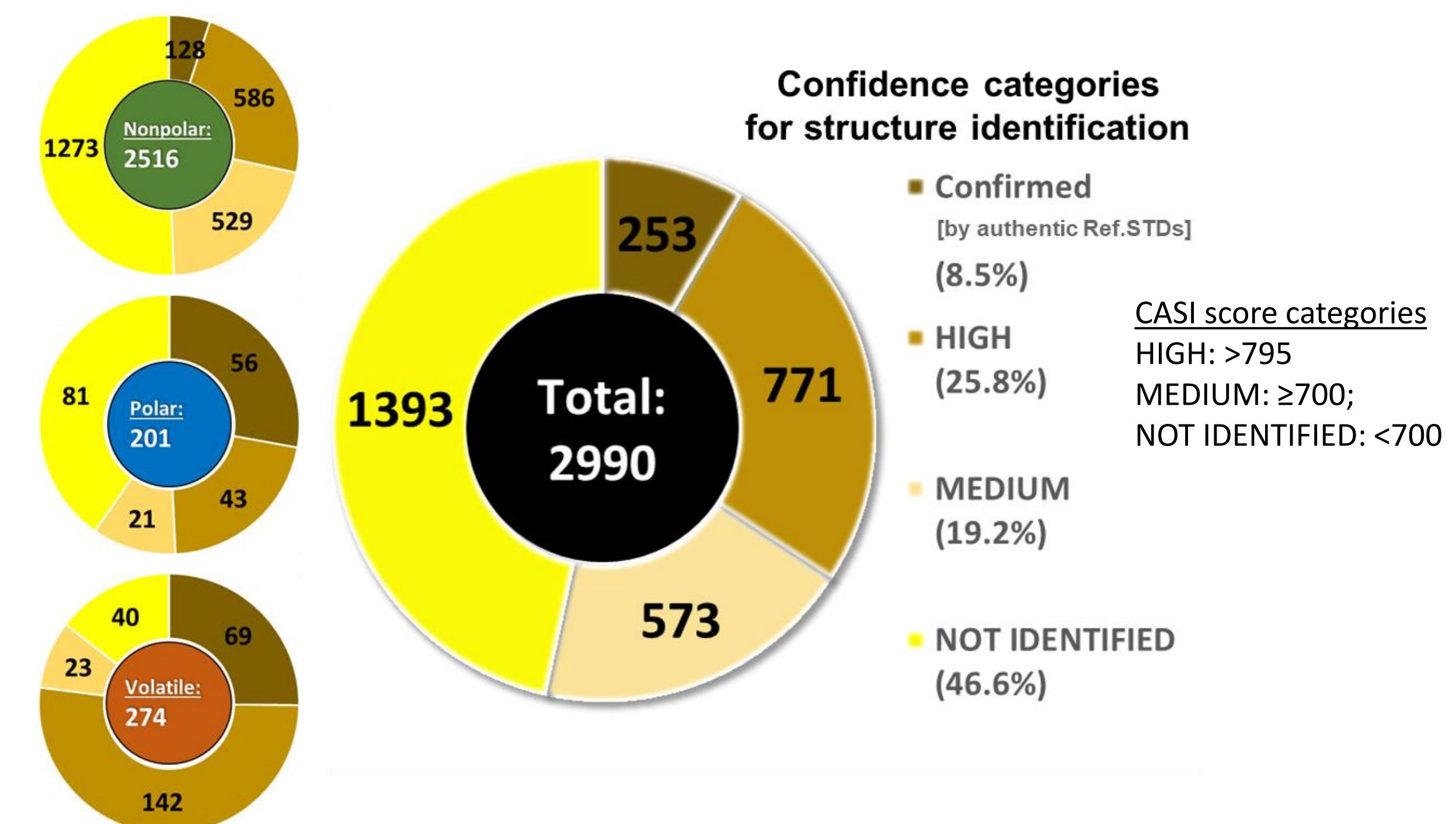
Non-targeted screening of 3R4F reference cigarette whole smoke by GCxGC-TOFMS

Coverage of chemical space



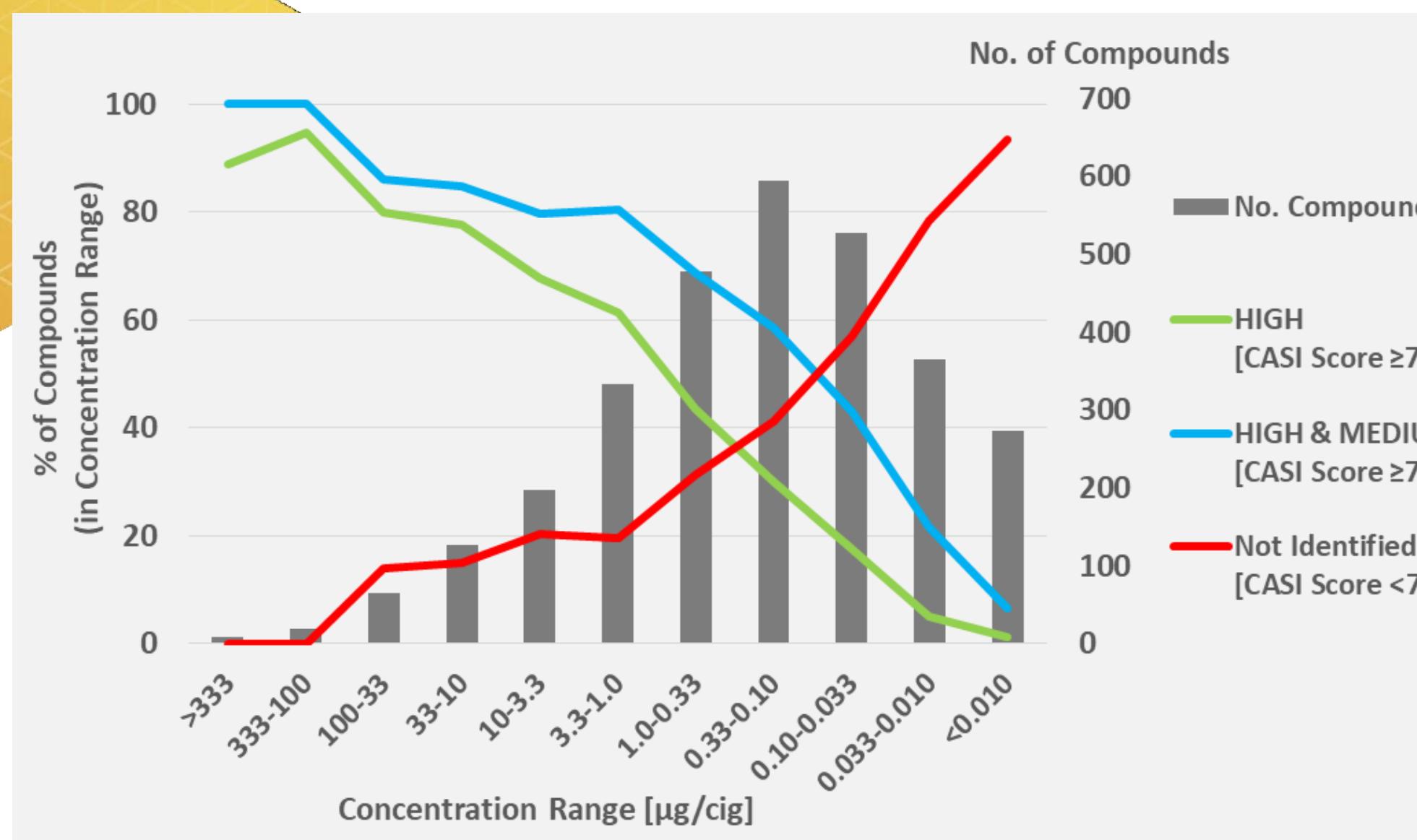
Confidence categories in Structure ID for 2990 compounds

found in 3R4F reference cigarette whole smoke by GCxGC-TOFMS



Performance CASI (GC — nominal mass)

Confidence CASI: Rate of meaningful proposals as a function of concentration in 3R4F reference cigarette smoke

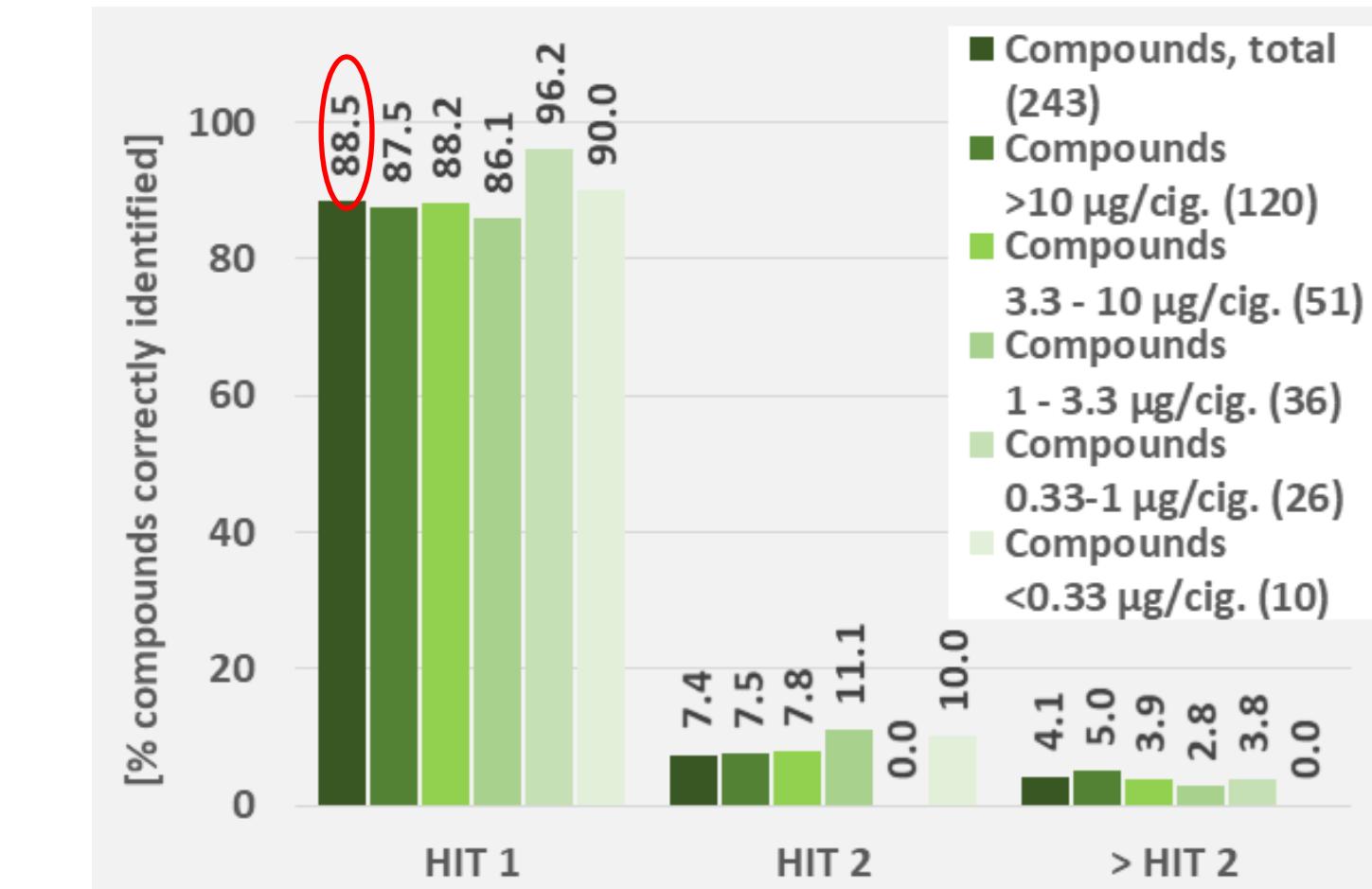


- Range >100 ng/cig.:**
Higher numbers of reliable structural proposals than unidentified structures
- Range <100 ng/cig.:**
 - Risk for lower spectral quality
 - Less-common structures, not available in commercial MS-DBs
(e.g., representative of unfavorable reaction chemistry, complex thermal degradation, or rearrangement processes)

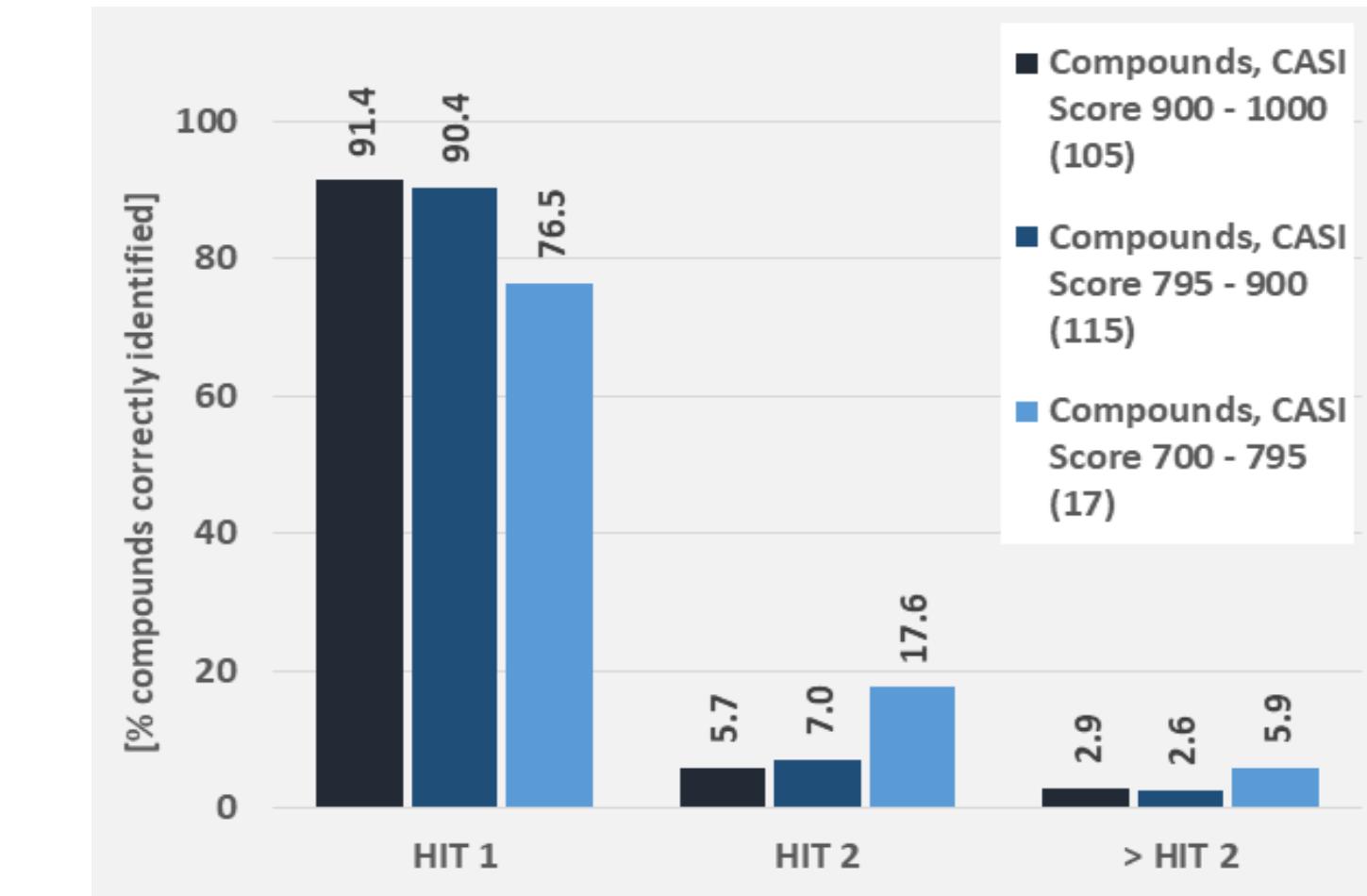
Knorr, A., Almstetter, M. et al., "Performance evaluation of a non-targeted platform using GC_xGC-TOFMS....", Analytical Chemistry, Jun 2019
<https://pubs.acs.org/doi/10.1021/acs.analchem.9b01659>

Confidence as a function of concentration

(confirmed by reference standards in 3R4F reference cigarette smoke dataset)



...as a function of CASI Score

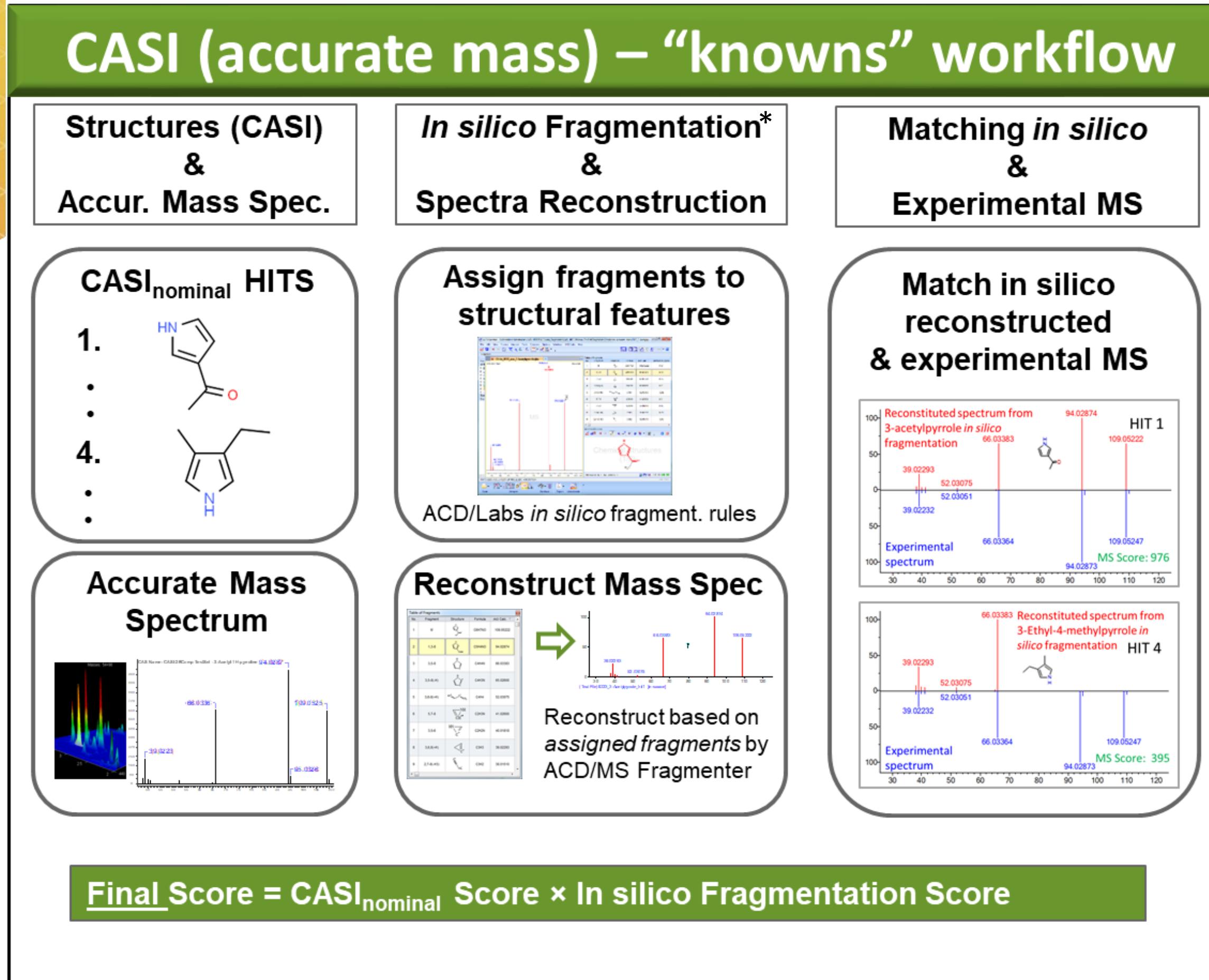


- Consistent good performance maintained down to concentration range <0.33 μg/cig.
- True positive rate (sensitivity): 88.5 %

High-throughput structure identification (GC — accurate mass)

CASI for GCxGC-HRAM-TOFMS — Compounds present in Mass Spectral (MS) databases (“knowns”)

- *In silico* assignment of fragments to structural features



*ACD/MS Fragmenter, version 2015.2.5, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2019.

- Proof-of-concept (PoC) test set (28 compounds)

- Subset of 28 confirmed compounds selected for proof-of-concept study (*out of >2,900 for 3R4F reference cigarette smoke above 100 ng/item*)
- Selection: Diversity and complexity of structures & low to high specific mass spectra

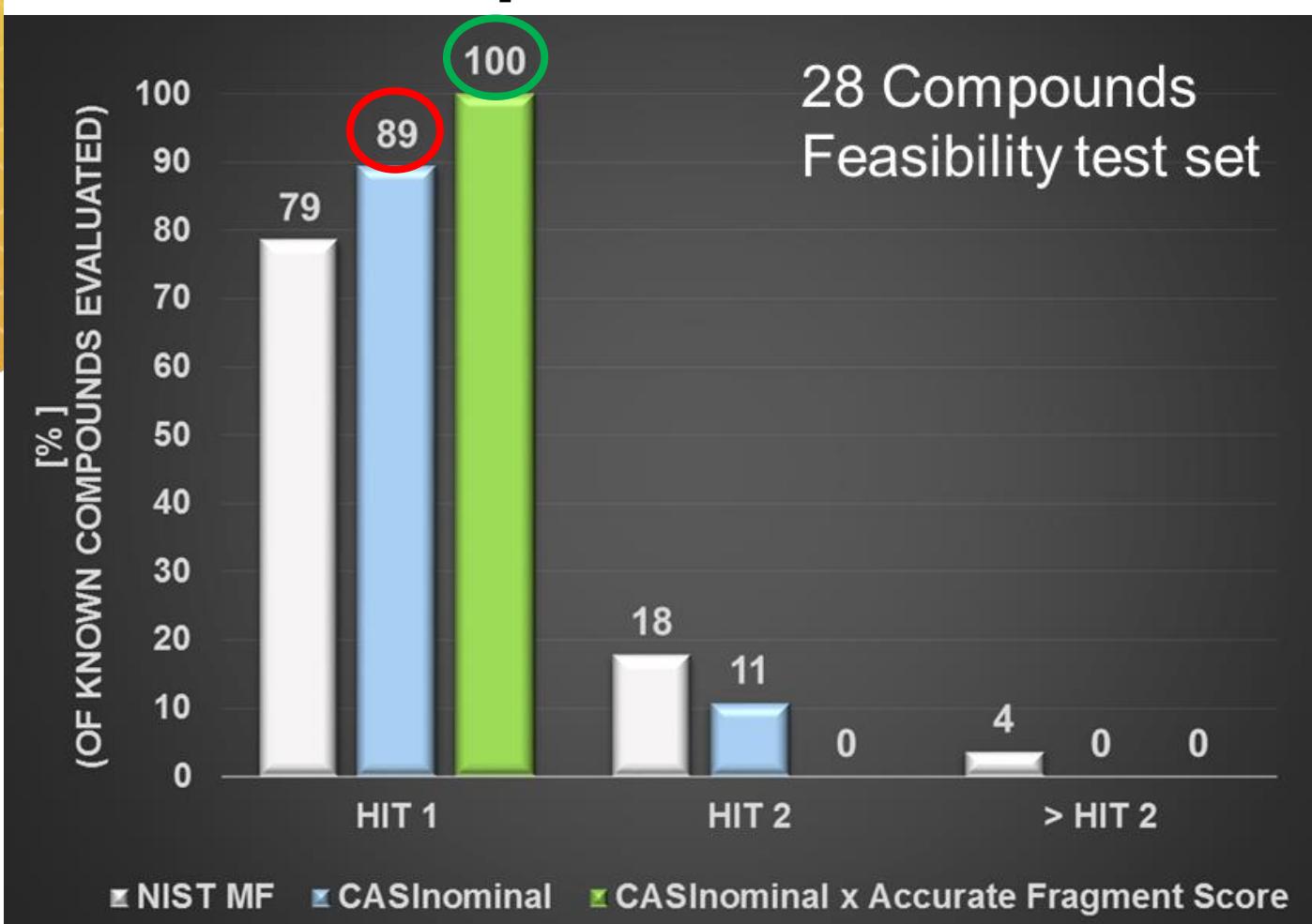
Compound Name	Structure	No.	Class	CAS	MW	FORMULA	NIST MF	CASI _{nom.} Score	CASI _{nom.} x acc.mass
Pyridine, 3-methyl-		1	pyridine, alkyl-	108-99-6	93.0578	C6H7N	950	937	903
2-Methylpyrazine		2	pyrazine, alkyl-	109-08-0	94.0531	C5H6N2	962	927	891
3-Pyridinol		3	pyridine, hydroxy-	109-00-2	95.0371	C5H5NO	832	788	743
2-Furanmethanol		4	furan, hydroxy-	98-00-0	98.0368	C5H6O2	935	899	809
2,3-Pentanedione		5	α-dicarbonyl	600-14-6	100.0524	C5H8O2	930	835	807
Methyl pyruvate		6	ester	600-22-6	102.0317	C4H6O3	952	778	735
Butanoic acid, 2-methyl-		7	small acid, branched-	116-53-0	102.0681	C5H10O2	873	848	846
3-Vinylpyridine		8	pyridine, vinyl-	1121-55-7	105.0578	C7H7N	925	905	856
p-Cresol		9	phenol	106-44-5	108.0575	C7H8O	933	932	695
2-Ethylpyrazine		10	pyrazine	13925-00-3	108.0687	C6H8N2	832	798	754
3-Acetylpyrrole		11	pyrrole, carbonyl-	1072-82-8	109.0528	C6H7NO	941	938	915
Cyclotene		12	cycl. ketone, hydroxy-	80-71-7	112.0524	C6H8O2	948	935	906
Phenylacetaldehyde		13	arom. aldehyde	122-78-1	120.0575	C8H8O	906	882	851
5-Hydroxymethylfurfural		14	furfural, hydroxy-	67-47-0	126.0317	C6H6O3	881	858	829
Methyl 2-furoate		15	furan, carboxylic ester	1334-76-5	126.0317	C6H6O3	923	843	538
Pyranone		16	pyranone, hydroxy	28564-83-2	144.0423	C6H8O4	904	800	799
2-Methoxy-4-vinylphenol		17	phenol, methoxy-	7786-61-0	150.0681	C9H10O2	911	902	863
Cotinine		18	alkaloid, oxy-	486-56-6	176.0950	C10H12N2O	935	864	833
Megastigmatrienone		19	cyclohexenone, alkenyl-	38818-55-2	190.1358	C13H18O	799	798	775
Scopoletin		20	cumarin, hydroxy-, methoxy-	92-61-5	192.0423	C10H8O4	900	842	758
(E)-Solanone		21	unsaturated keton	54868-48-3	194.1671	C13H22O	906	894	869
3-Hydroxy-β-damascone		22	cyclohexenol, alkyl-, carbonyl-	102488-09-5	208.1463	C13H20O2	796	752	704
Palmitic acid		23	sat. fatty acid	57-10-3	256.2402	C16H32O2	892	889	876
Farnesyl acetone		24	sesquiterpene keton, isoprenoid	1117-52-8	262.2297	C18H30O	868	796	781
Neophytadiene		25	diterpene, isoprenoid	504-96-1	278.2974	C20H38	912	909	897
Eicosane		26	aliph. hydrocarbon	112-95-8	282.3287	C20H42	890	889	871
t-Phytol		27	diterpen, hydroxy-, isoprenoid	150-86-7	296.3079	C20H40O	915	906	903
Vitamin E		28	chroman, isoprenoid	10191-41-0	430.3811	C29H50O2	875	874	777

- Range MW: 93 - 430

Performance CASI (GC — accurate mass — “knowns” workflow)

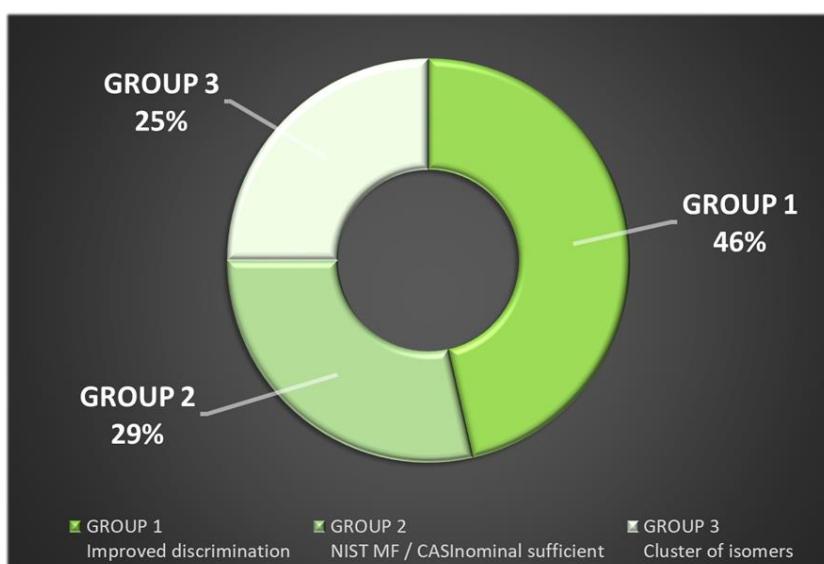
Performance PoC test set — “Knowns” workflow

- Correct structure ranked for HIT 1 for all compounds in test set



○ True positive rate CASI_{nominal}: 89 %
=> similar to previous 3R4F reference cigarette smoke set (88.5 %) => no bias

○ True positive rate CASI accurate mass: 100 %



- Group 1** (46 %) improved discrimination
- Group 2** (29 %) CASI_{nominal} sufficient
- Group 3** (25 %) Isomers cluster - no significant discrimination

True structure discriminated from other proposals, i.e. 3-acetylpyrrole

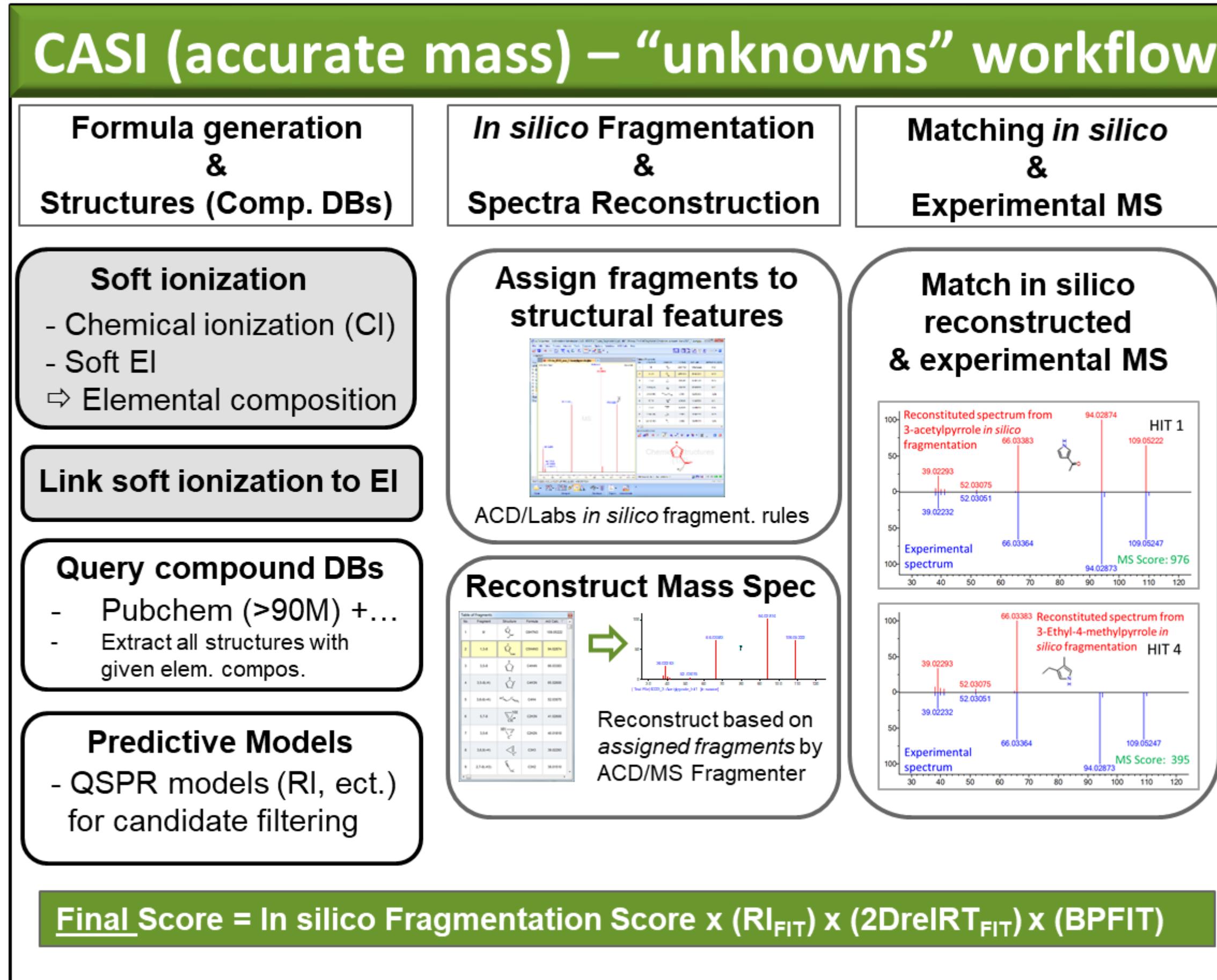


CASI HIT	Compound Name	Structure	NIST MF	CASI _{nom.} Score	MW	Formula	in silico Acc. Frag. Score	CASI _{nom.} x in silico Acc. Frag. Score
1	3-Acetylpyrrole (confirmed)	<chem>CC(=O)c1cc[nH]cn1</chem>	941	938	109.05276	C6H7NO	976	915
2	5-Bromo-3-methylidene-1-methoxy-cyclohexane	<chem>Oc1ccccc1Br</chem>	933	933	204.01498	C8H13BrO	37	35
3	Ethanone, 1-(1H-pyrrol-2-yl)-	<chem>CC(=O)c1cc[nH]cn1</chem>	897	879	109.05276	C6H7NO	976	858
4	3-Ethyl-4-methylpyrrole	<chem>CC(C)c1cc[nH]cn1</chem>	932	855	109.08915	C7H11N	395	338
5	1-Cyano-2-methylbuten-3-one	<chem>CC(=O)C(C)c1cc[nH]cn1</chem>	900	838	109.05276	C6H7NO	976	818
6	Pyridine, 3-methoxy-	<chem>Oc1ccccc1</chem>	725	721	109.05276	C6H7NO	976	704
7	Phenol	<chem>Oc1ccccc1</chem>	693	677	94.04186	C6H6O	37	25
8	3-Ethyl-2-methylpyrrole	<chem>CC(C)c1cc[nH]cn1</chem>	739	672	109.08915	C7H11N	395	266
9	3-Methylpyridazine	<chem>c1ccnnc1</chem>	663	645	94.05310	C5H6N2	395	255
10	2-Vinylfuran	<chem>CC=C1OC=CNC1</chem>	651	589	94.04186	C6H6O	37	22

High-throughput structure identification (GC — accurate mass)

CASI for GCxGC-HRAM-TOFMS — Compounds not present in MS databases (“unknowns”)

- *In silico* assignment of fragments to structural features



PoC test set (same 28 compounds + 1 additional compound; stigmasterol)

Formula generation not considered for PoC (concept under development)

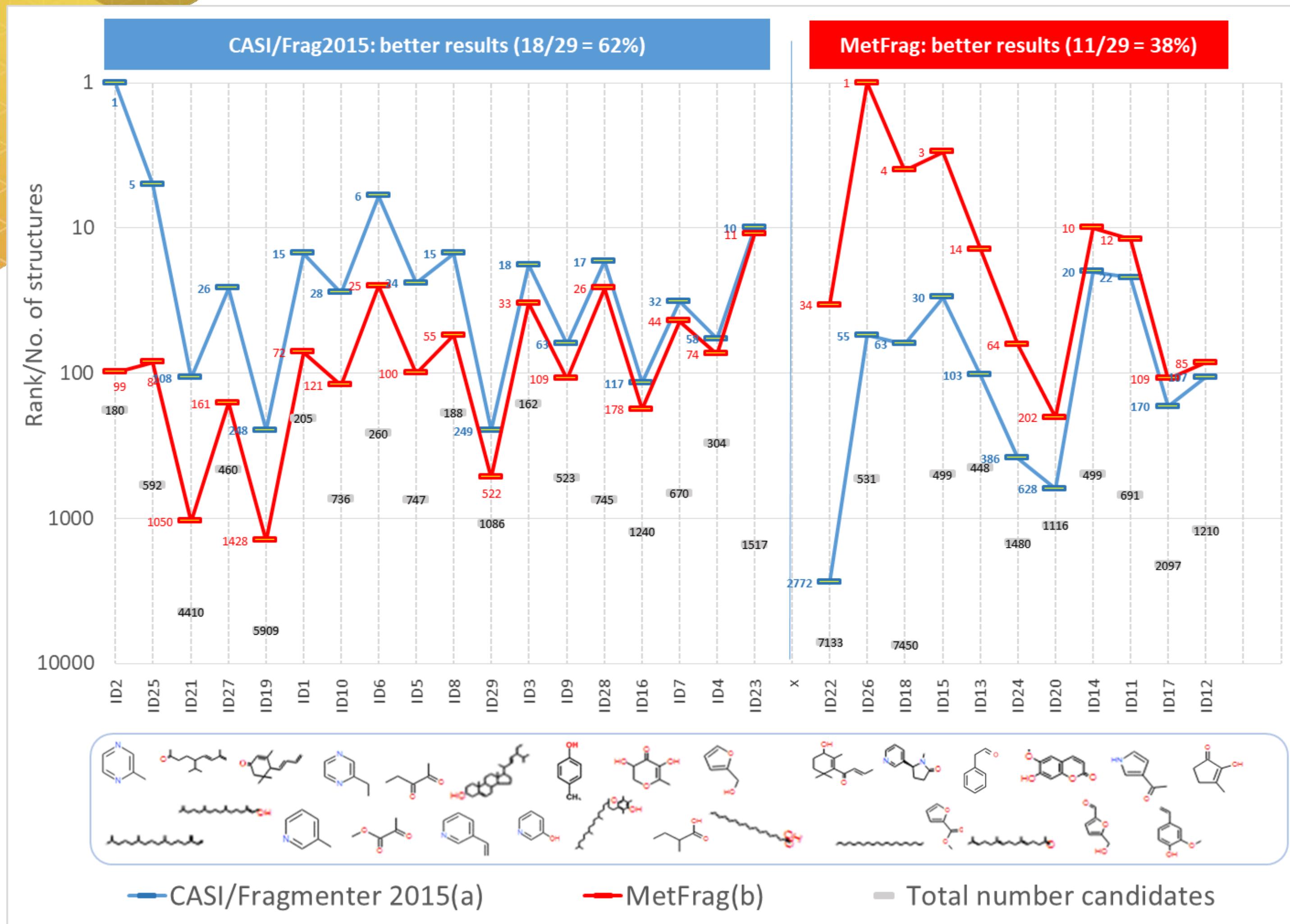
- Pubchem query for elemental composition of test candidates
⇒ 48K structures
- Remove radicals and conformational isomers
⇒ 35K structures
- Apply predictive models for RI and 2DreIRT and filter by using conservative cutoff limits
⇒ 32K structures ⇒ opportunity for improvement

Benchmark with MetFrag

- PoC test set (same 28 compounds)
- Structures derived from Pubchem (see above, including data curation)

High-throughput structure identification (GC — accurate mass)

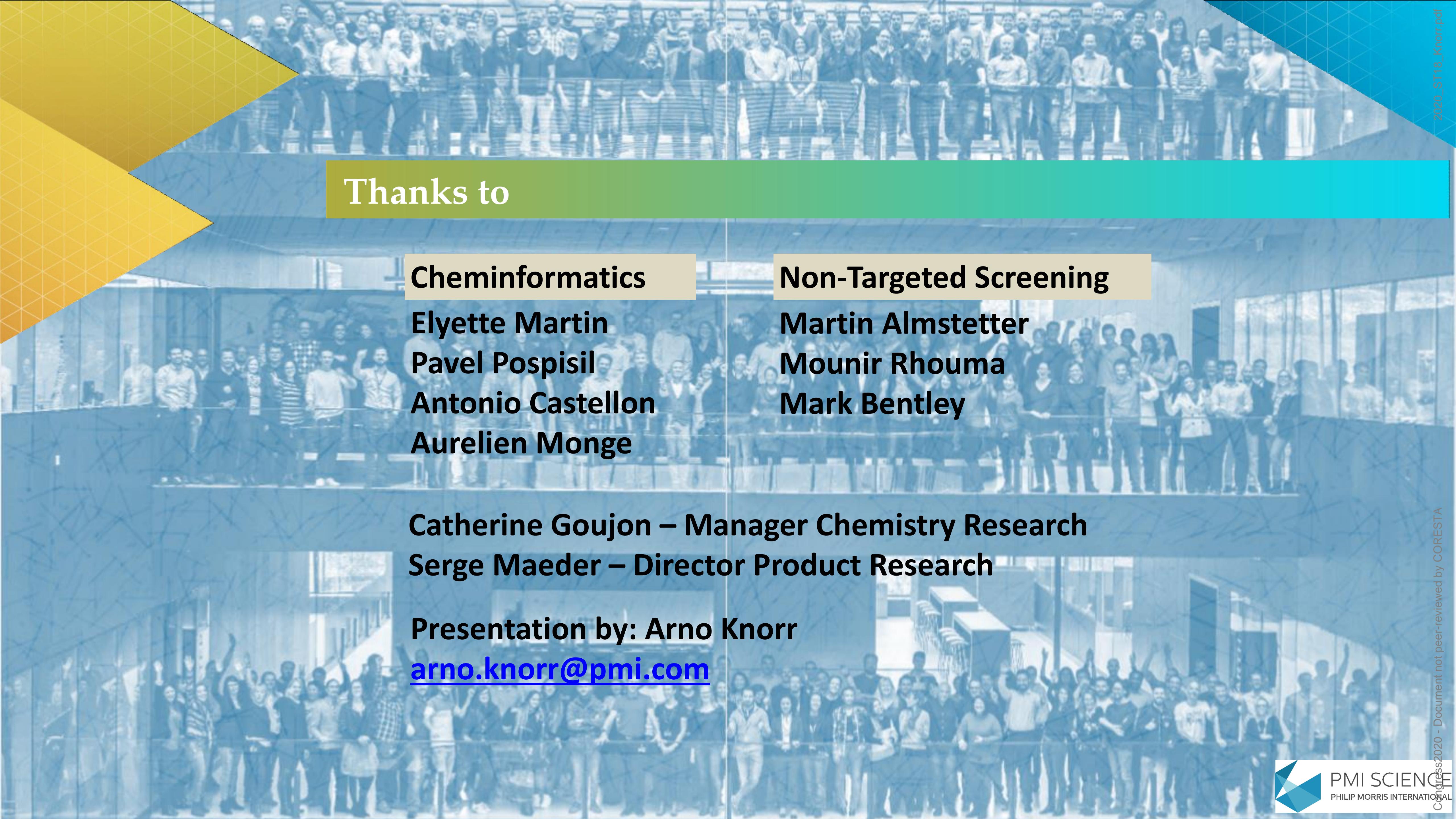
Results PoC test set – Compounds not present in MS databases (“unknowns”)



(a) ACD/MS Fragmenter, version 2015.2.5, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2019.

(b) MetFrag2.4.5-CL.jar

- Results for CASI and MetFrag are far below acceptable levels for automation
- CASI performed better than MetFrag in the PoC test set
(18/29 better ranked true hits ~62%)
- Next steps for improvement of CASI “unknowns” workflow:
 - Detailed gap assessment & key learnings
(e.g., *in silico* fragment assignment)
 - Test & optimize new ACD/MS Fragmenter, version 2018.1.1
 - Improvement on QSPR prediction models
(600 compounds in scope, actual model: 220)
 - Explore Bayesian models as an add-on
(e.g., compound class assignments)



Thanks to

Cheminformatics

Elyette Martin
Pavel Pospisil
Antonio Castellon
Aurelien Monge

Non-Targeted Screening

Martin Almstetter
Mounir Rhouma
Mark Bentley

Catherine Goujon – Manager Chemistry Research
Serge Maeder – Director Product Research

Presentation by: Arno Knorr
arno.knorr@pmi.com