

Normal boiling temperature estimation: Rarey-Nannoolal group contribution versus the GRAPPA graph neural network

Benchmark on the Rareytec Tb collection (10,026 compounds), and the resulting default-estimator policy of the PropertyBank platform

Rareytec Co., Ltd. – 2026-09-09

Summary

Two structure-based estimators of the normal boiling temperature T_b were run over the complete Rareytec T_b collection of 10,026 organic compounds: the Rarey-Nannoolal group-contribution method (RN) and GRAPPA, a graph neural network that predicts Antoine vapour-pressure parameters, from which T_b follows as the temperature where the vapour pressure equals 101.325 kPa. On the 8404 compounds both methods can treat, GRAPPA is the more accurate estimator for the typical molecule: its median absolute deviation is 5.2 K against 8.7 K for RN, and it is the closer of the two for 5013 compounds against 3381. RN, however, is considerably more robust for large molecules: from about 16 heavy (non-hydrogen) atoms upward its average deviation is smaller, and beyond 25 heavy atoms GRAPPA's average deviation is roughly twice that of RN. For the homologous series of n-alkanes GRAPPA diverges systematically above C_{26} (+25 K at C_{26} , +61 K at C_{32}) while RN stays within ± 7 K.

As a consequence the PropertyBank platform now uses GRAPPA as the preferred T_b and vapour-pressure estimator wherever it is applicable, with the exception of structures with 16 or more heavy atoms and of the families in which RN showed the smaller deviations (acid chlorides, phosphates and the RN catch-all family "All Other Compounds"). For those structures the Rarey-Nannoolal value stays the default and the GRAPPA value is displayed alongside for comparison. Where RN has no group coverage the GRAPPA value is used regardless, since an estimate is better than none. Both methods remain visible and selectable on every component page.

The comparison also flagged 889 reference values for which both methods deviate by more than 25 K in the same direction; these are being reviewed as suspect entries of the collection rather than as method failures.

1. Introduction

The normal boiling temperature is the single most important pure-component property for process simulation: it anchors the vapour-pressure curve, enters most critical-property and enthalpy-of-vaporisation estimates and decides the feasibility of a distillation. For the many compounds without a measured value an estimate from the molecular structure has to serve. The Rareytec platform has so far relied on one such estimator, and this report documents why a second one has now been added and how the two are combined.

Rarey-Nannoolal group contribution (RN). The method of Nannoolal, Rarey, Ramjugernath and Cordes (Fluid Phase Equilibria 226, 2004, 45-63) describes the molecule by first-order structural groups, second-order corrections for interactions between groups and a size-dependent normalisation, $T_b = \text{sum of group contributions} / (n^{0.6583} + 1.6868) + 84.3395$ K, with n the number of non-hydrogen atoms. The group table was regressed on a collection of experimental normal boiling temperatures assembled in 2004. Its strengths are transparency (every contribution can be inspected), a well-behaved extrapolation with molecular size, and a clear statement of applicability: where a group is missing the method declines rather than guessing. The Rareytec implementation is a Python re-implementation of the published method with the

group assignment performed by RDKit, extended by the associated vapour-pressure model of the same authors.

GRAPPA. GRAPPA (Hoffmann, Hasse and Jirasek, 2025; code published under the MIT licence at github.com/marco-hoffmann/GRAPPA) is a graph attention network (GATv2) that reads the molecular graph directly and outputs the three Antoine parameters A, B, C of $\ln(p_{\text{sat}}/\text{kPa}) = A - B/(T/K + C)$. The model was trained on experimental vapour-pressure data; the normal boiling temperature is not a training target but follows from the predicted curve as $T_b = B/(A - \ln 101.325) - C$. GRAPPA accepts neutral, single-fragment molecules containing at least one carbon atom and only the elements C, N, O, F, Cl, Br, I, S and P. Because the model is a black box, its output carries no group breakdown and no built-in warning when a structure lies outside the training domain; this benchmark is the platform's substitute for such a warning.

Why a benchmark on our own collection. Published accuracy figures of the two methods refer to different data sets and different subsets of chemistry. For a decision on which estimate to present as the platform default, both methods had to be scored on the same compounds with the same reference values. The Rareytec T_b collection (Section 2) is the natural test bed because it is exactly the population of compounds the platform users encounter.

Caveat: neither method is out-of-sample. RN was fitted on a 2004 boiling-point collection and GRAPPA on vapour-pressure data of largely the same well-measured compounds; the reference values used here overlap with both training sets to an unknown extent. The numbers in this report therefore describe how well each method reproduces the known boiling points of a broad organic collection, not a strict prediction test on unseen chemistry. This caveat applies equally to both methods and does not affect their relative ranking in the size and family breakdowns, which is what the platform policy is built on.

2. The T_b reference collection

The collection `tb_master.csv` (built 2026-07-24, frozen for this benchmark) holds one recommended normal boiling temperature per compound for 10,026 organic compounds, consolidated from public sources: the OPERA curated physchem data set (its high-confidence and its wider 2.9 release), PubChem experimental boiling-point annotations, Wikidata, and the NIST ThermoML archive (direct boiling points plus values interpolated from published vapour-pressure series at 101.325 kPa). Each entry carries the number of independent sources that agree within tolerance, a trust flag set by the consolidation rules, and the canonical SMILES from which both estimators work.

recommended source	compound s	agreeing sources	compounds
		1	7310
OPERA	5224	2	1914
OPERA29	2323	3	593
PubChem	1658	4	141
Wikidata	484	5 or more	68
ThermoML	267	trusted / untrusted	9835 / 191
ThermoML_VPinterp	70		

Table 1. Composition of the reference collection (all 10,026 compounds).

Values backed by two or more independent sources (2,716 compounds) are the most reliable references and are reported separately throughout. Single-source values, in particular from the wider OPERA 2.9 release and from PubChem annotations, contain a noticeable share of decomposition temperatures, reduced-pressure boiling points and mis-assigned structures; these affect both estimators identically and are discussed in Section 4.7.

3. Procedure

RN was evaluated with the platform's own engine. GRAPPA was evaluated on a local instance of the model served on the Rareytec lab server, so that no structure left the lab network; the instance reproduces the public GRAPPA release. For every compound both methods received the same canonical SMILES. A method "covers" a compound when it returns a value: RN declines when a required group is missing from its table, GRAPPA when the element or charge constraints above are violated. All statistics are computed on the deviation $d = T_{b,estimate} - T_{b,reference}$ in kelvin:

bias = mean of d ; AAD = mean of $|d|$ (average absolute deviation); medAD = median of $|d|$, which is insensitive to the few catastrophic outliers that dominate AAD and RMSE; RMSE = root mean square of d ; the fraction of compounds within 10 K and beyond 25 K; and the head-to-head count, i.e. for how many compounds each method is the closer of the two. Unless stated otherwise the tables refer to the common subset of 8404 compounds covered by both methods, so that every comparison is on identical references.

4. Results

4.1 Coverage

	RN	GRAPPA	both	RN only	GRAPPA only	neither
compounds covered	8918	9043	8404	514	639	469

Table 2. Coverage of the 10,026 compounds.

GRAPPA covers slightly more of the collection. The 639 compounds only GRAPPA can treat are mostly large or unusual structures for which RN lacks a group; GRAPPA's deviations there are poor (AAD 174 K, medAD 32 K) and are the reason the platform keeps offering a GRAPPA value where RN has none, but with the caveat that these values are extrapolations. The 514 compounds only RN can treat are mainly organosilicon compounds (about half) and structures written with formal charges, e.g. charge-separated nitro groups, which GRAPPA rejects; RN's AAD there is 30 K (medAD 10 K).

4.2 Overall accuracy on the common subset

subset / method	n	bias / K	AAD / K	medAD / K	RMSE / K	≤10 K	>25 K
RN, all 8918 covered	8918	+8.9	22.0	8.7	50.0	55%	18%
GRAPPA, all 9043 covered	9043	+28.0	35.5	5.7	189.1	65%	19%
RN, common subset	8404	+8.3	21.5	8.7	48.6	56%	18%
GRAPPA, common subset	8404	+17.8	25.0	5.2	102.1	68%	16%
RN, trusted references	8271	+8.5	20.5	8.5	47.5	56%	17%
GRAPPA, trusted references	8271	+15.9	22.2	5.1	93.6	69%	15%
RN, untrusted references	133	-1.8	78.5	72.9	93.0	8%	86%
GRAPPA, untrusted references	133	+133.6	197.7	133.1	338.0	3%	95%
RN, multi-source (≥2 sources)	2344	+1.9	13.1	6.5	26.4	65%	11%
GRAPPA, multi-source (≥2 sources)	2344	+5.9	10.4	3.1	27.7	79%	9%
RN, single-source	6060	+10.8	24.7	9.5	54.8	52%	20%
GRAPPA, single-source	6060	+22.3	30.6	6.4	119.0	63%	19%

Table 3. Deviation statistics. Head-to-head on the common subset: RN closer for 3381 compounds, GRAPPA closer for 5013, ties 10. Correlation of the two methods' deviations $r = 0.65$.

The two methods disagree on which statistic they win. GRAPPA has the far better median (5.2 against 8.7 K) and places 68% of the compounds within 10 K against 56% for RN; on the well-documented multi-source references its AAD of 10.4 K is also clearly below RN's 13.1 K. RN, on the other hand, has the smaller average and root-mean-square deviations on the full common subset, because GRAPPA produces a small number of extreme failures (several thousand kelvin for sulfonated azo dyes, Section 4.6) that RN, extrapolating linearly in the group contributions, does not. Both methods are biased high (estimates above the reference on average); this is

largely a property of the references, many of which are decomposition or reduced-pressure temperatures that lie below the true normal boiling point.

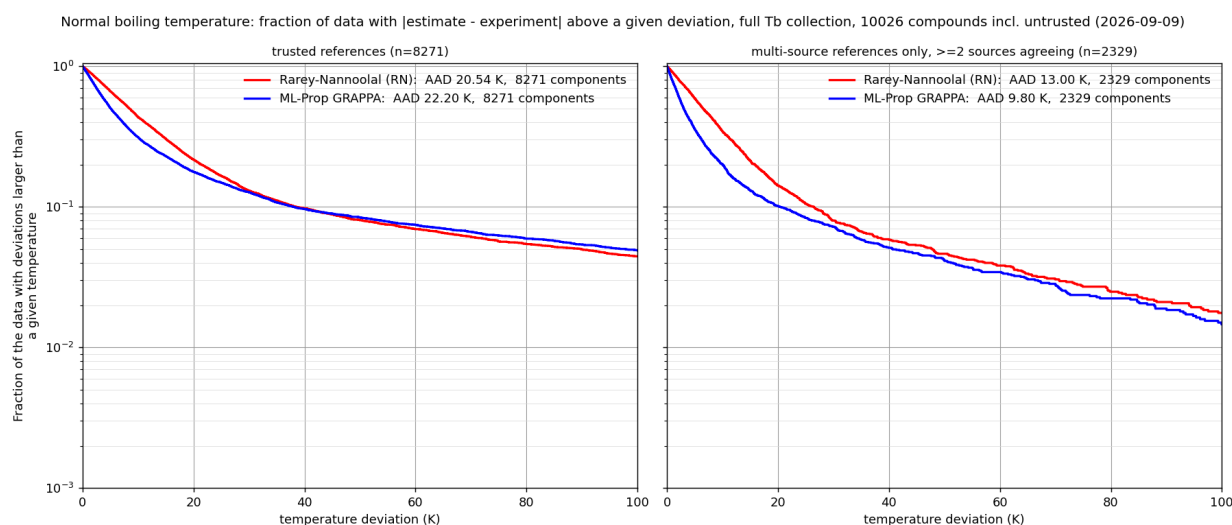


Figure 1. Fraction of compounds whose absolute deviation exceeds the value on the abscissa, for the trusted references (left) and the multi-source references (right). GRAPPA is better up to about 40 K; beyond that the curves cross and RN has fewer large failures.

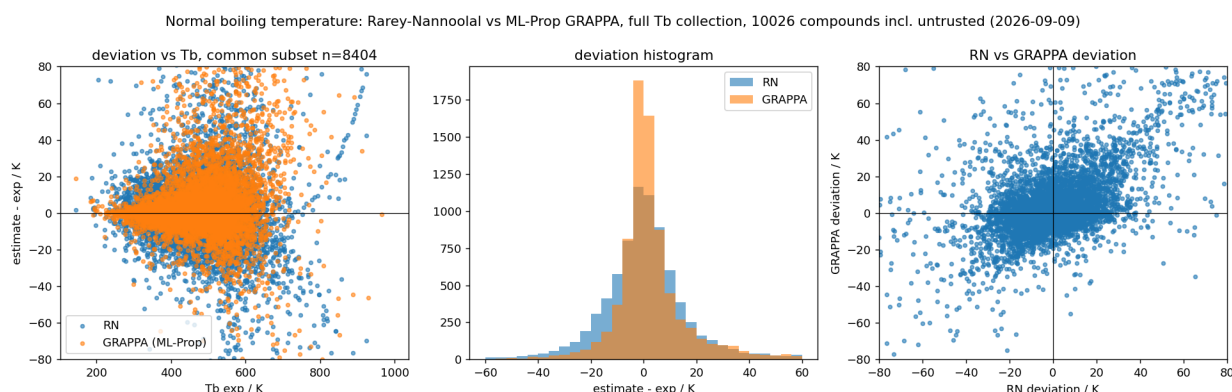


Figure 2. Common subset: deviation against the reference boiling temperature (left, clipped at ± 80 K), histogram of the deviations (centre) and the two methods' deviations against each other (right). The elongated cloud in the right panel and $r = 0.65$ show that a large part of both methods' error is shared, i.e. sits in the reference values.

4.3 Dependence on molecular size

heavy atoms	n	RN AAD	GRAPPA AAD	RN medAD	GRAPPA medAD	RN closer	GRAPPA closer
1-3	47	14.7	2.9	8.4	1.2	6	41
4-5	459	12.2	6.3	7.2	2.5	121	337
6-10	3820	12.7	8.9	6.9	3.8	1304	2513
11-15	2643	16.8	16.6	8.8	6.1	1129	1510
16-25	1096	40.9	50.3	17.5	17.1	586	508
26-40	185	110.9	198.1	45.1	72.9	122	63
41+	21	253.3	521.4	261.8	315.6	12	9

Table 4. Trusted references of the common subset, binned by the number of non-hydrogen atoms (K).

This is the decisive breakdown. For the smallest molecules (up to 5 heavy atoms) GRAPPA's advantage is largest: its median deviation is a third of RN's and it is closer in about three quarters of the cases. Between 6 and 10 heavy atoms, the most populated bin, GRAPPA is still clearly better on every statistic. Between 11 and 15 heavy atoms the averages are already level while GRAPPA keeps the better median and the head-to-head count. Between 16 and 25 heavy atoms the medians are equal but GRAPPA's average is the larger, and RN wins the head-to-head count; from 26 heavy atoms upward RN is better on all counts and the gap widens with size. The neural

network has seen few large molecules in training, and where it extrapolates it does so without the physical constraint that a group-contribution sum provides. The boundary of 16 heavy atoms is where the head-to-head count tips and is used as the platform's size rule.

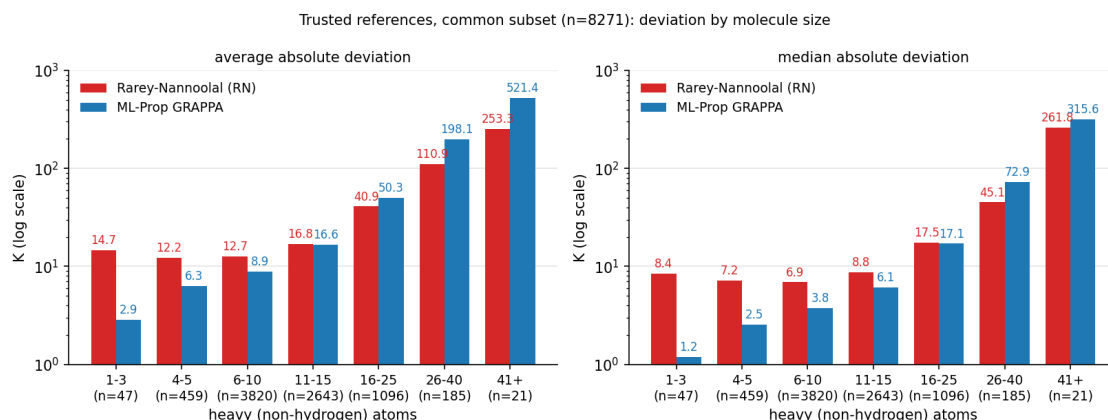


Figure 3. Average and median absolute deviation by molecule size, trusted references (logarithmic scale so that the small-molecule bins stay readable next to the large-molecule ones).

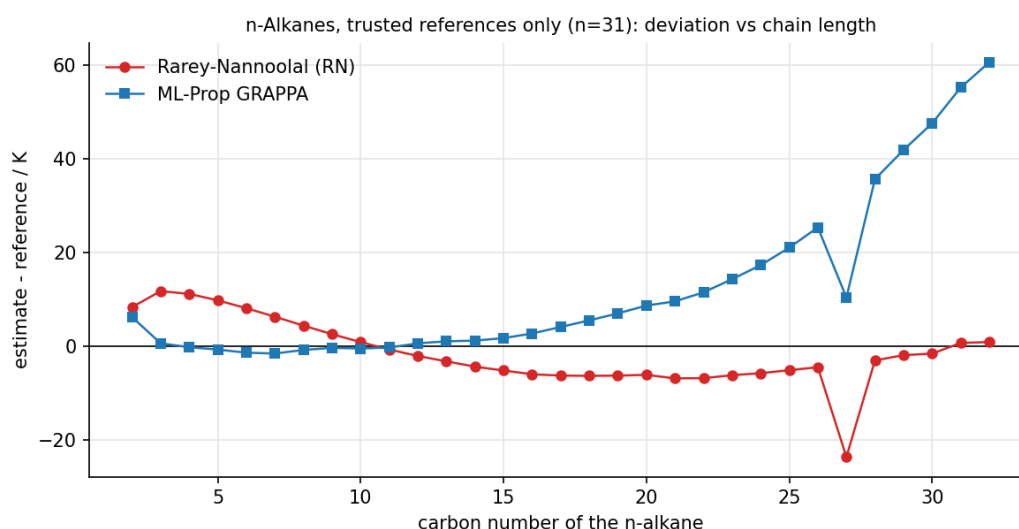


Figure 4. The n-alkane series (trusted references, ethane to dotriacontane). GRAPPA follows the references to within 2 K up to C_{15} , then drifts upward to +61 K at C_{32} ; RN shows its known slight overestimation of the short chains and stays within ± 7 K over the whole series (the C_{27} point is a reference-value outlier hit by both methods).

4.4 Dependence on the reference source

recommended source	n	RN bias	GRAPPA bias	RN AAD	GRAPPA AAD	RN medAD	GRAPPA medAD
OPERA	4708	+0.1	+3.7	11.1	8.5	6.6	3.8
OPERA29	1775	+18.2	+39.9	34.9	51.4	13.0	11.0
PubChem	1224	+24.1	+40.8	43.3	53.3	15.7	13.5
Wikidata	399	+17.2	+25.2	25.9	30.0	12.4	4.7
ThermoML	241	+2.0	+3.7	9.8	5.7	5.8	1.4
ThermoML_VPinterp	57	-0.9	+0.1	9.7	5.7	8.4	3.2

Table 5. Common subset by the source of the recommended reference value (K).

The quality of the reference matters more than the choice of method. On the curated sources (OPERA high confidence, ThermoML) both methods agree with the data to within about 10 K on average and GRAPPA to within a few kelvin in the median; on the less curated sources (OPERA 2.9, PubChem annotations) the deviations of both methods grow three- to four-fold and the bias turns strongly positive, consistent with a population of reference values that are not true normal boiling points.

4.5 Chemical families

Every compound is assigned to the nested RN families of its structural groups (a compound can belong to several, e.g. Hydrocarbons > Saturated HC > n-Alkanes), and the statistics are recomputed per family. Figure 5 shows all families with at least 20 common compounds; Table 6 gives an excerpt. GRAPPA is better in most functional-group families, typically by a factor of two in the median. RN is better for hydrocarbons in general and for saturated and aromatic hydrocarbons and n-alkanes in particular, for acid chlorides and phosphates, and for the catch-all family "All Other Compounds". For amines, amides, nitriles and most other nitrogen families GRAPPA is the better choice.

family	n	RN AAD	GRAPPA AAD	RN medAD	GRAPPA medAD	RN : GRAPPA closer
n-Alkanes	85	48.8	90.8	23.6	118.2	67 : 18
Alkanes (non-cyclic)	237	21.4	36.4	5.0	5.0	145 : 91
Saturated HC	366	16.4	25.4	5.0	4.4	194 : 171
Aromatic HC	291	10.2	16.4	6.3	3.6	112 : 178
Hydrocarbons (HC)	1004	12.7	16.2	5.8	3.6	395 : 607
Acid Chloride	32	7.7	9.8	2.8	5.6	23 : 9
Phosphates	20	35.9	46.8	19.3	18.9	14 : 6
All Other Compounds	34	73.0	92.5	25.4	30.1	24 : 10
Lactones (HC)	60	27.3	27.3	15.1	8.8	27 : 33
Alcohols (HC)	490	11.6	9.4	7.0	3.4	165 : 323
Esters (HC, no ==O)	612	12.8	13.0	5.5	4.9	305 : 305
Amines	927	37.2	60.6	10.7	7.4	454 : 473
Halogenated Compounds	2110	18.3	16.9	9.2	4.2	691 : 1417
Other Oxygenated Compounds	5283	21.5	25.0	8.6	6.2	2338 : 2938
Anhydrides Chains	25	48.3	17.1	33.1	9.0	3 : 22
Carbamate	21	54.2	36.7	48.8	6.4	6 : 15
Di Amides	46	20.5	13.4	16.4	7.5	10 : 36

Table 6. Excerpt of the family statistics, common subset (K). The complete table for 79 families is in the accompanying CSV file.

Most of the family effects are a proxy for molecular size: the families in which RN wins are those populated by long chains and large molecules (n-alkanes, phosphate triesters, the residual class), and within a family the crossover with size seen in Section 4.3 recurs. The size rule therefore captures most of what a family rule would, and only the three families in which RN remains better after the size split are carried into the platform policy.



Figure 5. Average (left) and median (right) absolute deviation per chemical family, common subset, families with at least 20 compounds. Values beyond the axis range are annotated.

4.6 Largest deviations

compound	reference / K	RN dev. / K	GRAPPA dev. / K	source
URUSHIOL	473	+825	n/a	PubChem
Dipentaerythritol hexastearate	623	+621	+353	OPERA29
Streptomycin	348	+578	+4435	PubChem
Pantoprazole	333	+516	+548	PubChem
Iscotrizinol	513	+498	+2486	OPERA29
Daunorubicin	463	+462	+2509	PubChem
((2,2',2''-(Nitrilotris(2,1-ethanediy)(nitrilo-kn)methylidyn	383	+452	+824	OPERA29
2,2-Bis[4-[bis(phenoxy)phosphoryloxy]phenyl]propane	474	+446	+517	OPERA29

Table 7. Largest RN deviations.

compound	reference / K	GRAPPA dev. / K	RN dev. / K	source
1,4-Benzenedisulfonic acid, 2,2'-[1,2-ethenediyl]bis[(3-sulfo	623	+6470	n/a	OPERA29
Hexasodium 2,2'-(vinylenebis((3-sulphonato-4,1-phenylene)imi	383	+6016	n/a	OPERA29
Tetrasodium 4,4'-bis(4-(4-sulfophenylamino)-6-(bis(2-hydroxy	623	+5623	n/a	OPERA29
1,5-Naphthalenedisulfonic acid, 2-[2-[8-amino-7-[2-[5-[4-ch	573	+4459	n/a	OPERA29
Streptomycin	348	+4435	+578	PubChem
CAS 2138848-94-7	543	+4342	n/a	OPERA29
2-Naphthalenesulfonic acid, 5,5'-[(1-methylethylidene)bis(4,	623	+3524	n/a	OPERA29
Trisodium 7-((4-((4-((2-(2-(ethenylsulfonyl)ethoxy)ethyl	673	+3468	n/a	OPERA29

Table 8. Largest GRAPPA deviations.

The worst cases of both methods are not boiling points at all: they are decomposition or melting temperatures of natural products, drugs and dyes entered as boiling points, and multiply sulfonated or charged dye structures far outside both methods' domains. Their character differs, though. RN's worst errors are a few hundred kelvin, bounded by the additive group sum; GRAPPA's reach several thousand kelvin because a small change of the predicted Antoine parameters moves the 101.325 kPa intersection to absurd temperatures. This is the second reason for the size rule: a group-contribution failure is visible and bounded, a neural-network failure is not.

4.7 Suspect reference values

For 889 compounds both methods deviate by more than 25 K in the same direction. Two independent estimators agreeing against the reference is strong evidence that the reference, not the estimate, is wrong. A few examples: hexabromocyclododecane (reference 463 K, both methods about +90 K: the value is a decomposition point), 1,1,1,2-tetrabromoethane (473 K, both +40 K), decabromodiphenyl ether (803 K, both +30 to +75 K). The list is stored as `tb_master10026_suspect_refs.csv` and feeds the next review of the collection; the review will re-check these entries against their original sources and, where a value is confirmed as a decomposition or reduced-pressure temperature, mark it as such rather than as a normal boiling point. Until then these entries keep their present trust flags, so the statistics above are, if anything, pessimistic for both methods.

5. Consequences for the PropertyBank platform

On the basis of these results the platform's automatic estimates for the normal boiling temperature and the vapour-pressure curve follow this policy from platform version 3.004.142 on:

1. GRAPPA is the **preferred** estimator whenever it is applicable to the structure (neutral, single-fragment, carbon-containing molecule of the elements C, N, O, F, Cl, Br, I, S, P) and the local GRAPPA service is reachable.
2. **Exception 1, size:** for structures with 16 or more heavy atoms the Rarey-Nannoolal value is the default.
3. **Exception 2, families:** for structures that RN assigns to the families Acid Chloride, Phosphates or "All Other Compounds" the Rarey-Nannoolal value is the default.
4. Under both exceptions the GRAPPA value is still computed and shown alongside as the comparison, with the reason for the preference spelled out, and the user can select either value as the working estimate.
5. The exceptions apply only where RN actually yields a value. Where RN has no group coverage for the structure, GRAPPA is used regardless, as an estimate is better than none; such values for

large molecules should be treated as extrapolations.

6. Where GRAPPA is not applicable or the service is unavailable, the Rarey-Nannoolal value is used as before.

7. The vapour-pressure curve follows the same choice: GRAPPA's predicted Antoine parameters are used verbatim when GRAPPA is preferred; otherwise the RN vapour-pressure model is fitted, with the GRAPPA curve as fallback when RN cannot provide one.

The policy will be revisited when the suspect references have been reviewed and when either method is updated. All benchmark scripts and result files are kept with the collection so that the comparison can be re-run unchanged.

6. Files

file	content
tb_master.csv	the reference collection (frozen)
mlprop_tb.py, compare_rn_mlprop.py	GRAPPA batch client and the RN / GRAPPA comparison (statistics, figures 1-2, suspect list)
report_figs.py, family_compare.py	size and n-alkane figures; per-family statistics and figure 5
tb_master10026_compare.csv	per-compound reference, both estimates, deviations, families (10,026 rows)
tb_master10026_families.csv	per-family statistics (79 families)
tb_master10026_suspect_refs.csv	889 references both methods contradict by >25 K
make_report.py	this report

References

Y. Nannoolal, J. Rarey, D. Ramjugernath, W. Cordes: Estimation of pure component properties. Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions. Fluid Phase Equilibria 226 (2004) 45-63.

Y. Nannoolal, J. Rarey, D. Ramjugernath: Estimation of pure component properties. Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions. Fluid Phase Equilibria 269 (2008) 117-133.

M. Hoffmann, H. Hasse, F. Jirasek: GRAPPA - a hybrid graph neural network for predicting pure component vapor pressures (2025). Code: github.com/marco-hoffmann/GRAPPA (MIT licence).