Physics of Fluids

A combined clustering/symbolic regression framework for fluid property prediction

Cite as: Physics of Fluids (in press) (2022); https://doi.org/10.1063/5.0096669
Submitted: 20 April 2022 • Accepted: 19 May 2022 • Accepted Manuscript Online: 21 May 2022

Filippos Sofos, Avraam Charakopoulos, Konstantinos Papastamatiou, et al.

Accepted to Phys. Fluids 10.1063/5.0096669

A combined clustering/symbolic regression framework for fluid property prediction

Filippos Sofos[1,a)], Avraam Charakopoulos[1], Konstantinos Papastamatiou[1], Theodoros E. Karakasidis[1]

AFFILIATION

[1]Condensed Matter Physics Laboratory, Department of Physics, University of Thessaly, 35100 Lamia, Greece.

[a)]Author to whom correspondence should be addressed: fsofos@uth.gr

ABSTRACT

Symbolic regression (SR) techniques are constantly gaining ground in materials informatics, as the machine learning counterpart capable of providing analytical equations exclusively derived from data. When the feature space is unknown, unsupervised learning is incorporated to discover and explore hidden connections between data points and may suggest a regional solution, specific for a group of data. In this work, we develop a Lennard-Jones (LJ) fluid descriptor based on density and temperature values and investigate the similarity between data corresponding to diffusion coefficients. Descriptions are linked with aid of clustering algorithms which lead to fluid groups with similar behavior, bound to physical laws. Keeping in mind that the fluid data space goes over the gas, liquid, and supercritical states, we compare clustering results to this categorization, and found that the proposed methods can detect the gas and liquid states, while distinct supercritical region characteristics are discovered, where fluid density and temperature affect the diffusion coefficient in a more complex way. The incorporation of symbolic regression algorithms on each cluster provides an in-depth investigation on fluid behavior, and regional expressions are proposed.

I. INTRODUCTION

It is maximally beneficial for science and engineering to exploit the power of artificial intelligence (AI) and machine learning (ML) for predicting fluid behavior and calculate properties in various timescales, simple and complex geometries, under ambient or extreme conditions[1]. Data science has a pivotal role in the whole process. Data from experiments and simulations, graphical data, and various literature data[2], are nowadays stored and being accessed from various electronic databases (DBs), in various formats. As their source may differ, it is of outmost importance to select and weight experimental and simulated datasets as a prerequisite step in the development of property models[3], and the detailed investigation for structure, patterns, and functional relationships in the data.

ML techniques have been incorporated to speed up the research procedure, by predicting properties that are time-consuming to extract or difficult to obtain experimentally, or, on the other hand, speeding up atomic-scale simulations with aid of ML-extracted potential maps. Common schemes for ML and simulations have created surrogate models able to bridge between different scales[4]. Research has shown that the construction of ML interatomic potentials (MLIPs) trained over ab-initio molecular dynamics (AIMD) output could enable the design of an efficient first-principles multiscale modeling, branching DFT with classical molecular dynamics (MD) and finite element (FE) simulations[5,6]. Such an achievement may lead to highly accurate property calculations, without paying the computational cost of sophisticated first principles calculations[7].

Furthermore, ML has been used in conjunction with experiments to accelerate material discovery[8], and even when data are scarce, synthetic data may be generated to complement the training model [9]. The ML training part is usually the most computationally expensive step, however, time spent may be regarded as an upfront investment of computational effort[10]. For complex fluid flows, Deep Neural Networks (DNNs) are often employed and being investigated on their ability to perform unsteady fluid dynamics and complex-shaped domain predictions[11]. DNNs are able to capture complex flow features through high-resolution image investigation, while running on parallel hardware architectures[12]. For simpler, faster and easier to implement models, such as Linear

Regression and Neural Networks (NNs), good performance on predicting turbulence flow properties is also reported[13].

A step further, learning from data in fluid mechanics has acquired strong physical basis since the incorporation of Symbolic Regression (SR), a technique that extracts a symbolic expression that matches data from an unknown function[14], by incorporating mathematical operators as building blocks to analytically describe physical phenomena. Equations proposed are not selected a priori; they are based on data and can be generalizable as long as they can be described in physical terms, with minimum prediction error and maximum simplicity[15]. This is particularly applicable in physics and material science, as most of the physical laws, when expressed as equations, are relatively mathematically simple[16] and differs from other regression-based methods where the function form is not a subject of investigation and only the numerical coefficients are optimized through fitting. Notwithstanding the fact that classical ML methods achieve high prediction and accuracy metrics[17], they are far from being able to be used as a practical tool to reveal and calculate desired properties[18] and SR, as continuously being evolved by new algorithms and tools[19,20], may be the solution.

ML methods usually entail supervised and unsupervised learning methods, where supervised learning refers to finding predictions for labeled data, while unsupervised learning is used for unlabeled data, without the need for labelling and supervision[21]. Supervised methods assume prior knowledge of the system's state and are limited to learning only known phases. For the unsupervised learning scenario, learning is achieved from the data itself without the need of prior labeling[22,23]. Features are extracted through an initial data pre-processing stage[24] and clustering[25,26] plays a vital role here, as a similarity measure to group data around a desired metric.

Data may come from various sources and MD simulations have traditionally been a continuous source for ML applications. During an MD simulation, interactions between particles (atoms or molecules) are calculated and this seems to be the most accurate method to investigate fluid behavior at the nanoscale[27–29]. To this end, Lennard-Jones model systems have been widely

investigated and results can be extrapolated to real fluids[30]. The radial distribution function (RDF) calculation can be the intermediate step towards extraction of the thermodynamic properties of the Lennard-Jones (LJ) fluid[31]. A back-propagation NN is utilized to construct an efficient mapping between the model parameters and four crucial physical properties of water, including the density, vaporization enthalpy, self-diffusion coefficient and viscosity[32], while the slip length at the nanoscale has been accurately predicted through Random Forest and NN methods[33].

In this work, a LJ fluid simulation dataset with diffusion coefficient data along a wide phase space, constructed by density and temperature conditions, is employed[34,35]. Prior knowledge of various fluid states (i.e., gas, liquid, supercritical) can be assumed from data[36], however, we consider an unsupervised learning approach and prompt the ML processes to discover data connections through clustering[37,38]. While fluid behavior across gas and liquid states is well-defined, behavior on the supercritical (SC) state and near the critical region, is somehow ambiguous. When compared to popular empirical relations, our results have shown that common empirical methods are inadequate to reproduce the complex behavior of the diffusion coefficient throughout the whole phase space.

SR is employed to propose analytical equations for each given cluster and discussion is made on the physical meaning of each equation. We believe that data science and statistical machine learning techniques inferred here can be valuable tools towards the full exploitation of the vast amount of data coming from fluid mechanics experiments and simulations, in order to substitute cases of expensive or extreme cases (e.g., extra high temperature or pressure experiments), complement their applicability, and extend output data post-processing and analysis.

II. THEORETICAL CONSIDERATIONS

A. Diffusion in SC fluids

There have been proposed various empirical relations in the literature to calculate the diffusion coefficients. Starting from the Chapman and Cowling (CC) equation[39] for diffusion in the low density regime, where particles are modeled as hard-spheres (HS), we obtain

$$D_{CC}^{*} = \frac{3}{8\sigma^2\rho^{*}}\sqrt{\frac{k_B T^{*}}{m\pi}} \tag{1}$$

For denser fluids, the CC equation has been adjusted by Speedy as[40]

$$D_{SP}^{*} = D_{CC}^{*}\left(1-\frac{\rho^{*}}{1.09}\right)\left(1+(\rho^{*})^2(0.4-0.83(\rho^{*})^2)\right) \tag{2}$$

Based on the above formulation, Zhu et. al[30] have proposed a diffusion equation that aims to cover the whole LJ fluid space, i.e., gas, liquid and SC, as

$$D_{Z}^{*} = \frac{3}{8\rho^{*}}\sqrt{\frac{T^{*}}{\pi}}\times A\times B \tag{3}$$

where, $A=\left(1-\frac{\rho^{*}}{\alpha(T^{*})^{b}}\right)\left[1+(\rho^{*})^{c}\left(\frac{P1(\rho^{*}-1)}{P2(\rho^{*}-1)+(T^{*})^{P3+P4\rho^{*}}}\right)\right]$ and $B=\exp\left(-\frac{\rho^{*}}{2T^{*}}\right)$, and $a, b, c, P1, P2, P3, P4$ a set of parameters[30].

However, Eqs. (1-3) may fail to provide accurate values for the diffusion coefficient inside the SC region, as transitions in the fluid's state from gas-like to liquid-like and vice versa may happen there. Significant changes in fluid density may happen over relatively small temperature ranges[41] and these changes are expected to correlate with other fluid properties above the critical temperature and density, and some of them are connected to the Widom line, i.e., the line that connects local maxima of thermodynamic properties[42]. During the last decades, SC fluids have gained increasing research and applied interest due to their principal role in fields like food, drug, chemical and energy industries[43]. Transport properties data, such as the diffusion coefficients for pure or binary

compounds, are needed for applications concerning mass transfer, which is a major field in separation and purification processes[44].

Understanding fundamental thermodynamic properties of the SC fluid is important. Inherent SC properties might include a combination of liquid-like density, and gas-like viscosity, compressibility and diffusivity, under high pressures[45]. Moreover, fluid properties can be adjusted closer to liquid or gas by varying pressure and temperature[46]. Research efforts have shown that deviations are expected from theoretical investigation on simple fluids based on the Chapman–Enskog (CE) approximate solution to the Boltzmann equation or the free volume theory[47], when it comes to a system of molecular species. To this end, the mean force kinetic theory (MFT) has been proposed for the calculation of the self-diffusivity of CO_2 in the SC region[48], while, an empirical relation that encompasses LJ fluid data and extrapolates to real substances has also achieved wide applicability in the field[30].

In such a multivariant environment, the accurate estimation of diffusion coefficients across all fluid states is important and statistical data science and ML techniques can be valuable tools towards this direction, if, furthermore, we take into account that classical macroscopic approaches fail to capture the detail needed to investigate the microscopic liquid/gas interface[49].

B. General Model Description

The dataset employed in this work has been derived from MD simulations data taken from 17 literature sources and consists of 927 points[34]. MD involves atomic interactions described by the potential u_{LJ}, which is a function of the distance r between the particles, as $u_{LJ} = 4\varepsilon[(\sigma/r)^{12} - (\sigma/r)^{6}]$, where σ and ε are the LJ parameters. The different number of simulated particles, the potential cut-off radius and different ensembles and thermostats employed in each work may induce some uncertainties. To eliminate such phenomena, data points are chosen in the basis of keeping the

cut-off radius close to 2.5σ and incorporating only systems with the largest number of particles to account for periodic boundary effects.

A general model of the steps followed in this work is depicted in Fig. 1a. We have performed clustering techniques in order to separate the diffusion coefficient data for the LJ fluid taken from our database[35] to groups of almost the same behavior. Two different algorithms have been tested: k-medoids and agglomerative hierarchical clustering[50]. We expect that this process can divide the original dataset in separate groups that correspond to different fluid states (gas, fluid, SC). Each cluster extracted is treated as a distinct dataset and is fed into the respective SR computational cell. The cells function in parallel and output mathematical expressions without any prior knowledge of their form, driven only by data. The first clean-up is made on the selection of equations that present low Mean Squared Error (MSE) and complexity (Comp.). The decision step follows, where physics-based explanation is performed on equations that adhere to these criteria. Each model stage is further analyzed in the next Sections.

FIG. 1. a) The symbolic regression data flow. Main steps are the clustering step, the SR computational step, and the physics-based decision step that finally selects the symbolic expressions that fit well to input data. b) An example of the tree-based SR process. A wide space of operators, constants and input variables create a pool from which random trees are formed in a parent-and-child manner. Only those with low MSE and Comp. are kept.

C. Clustering

Two popular clustering methods, the k-medoids and hierarchical clustering, have been exploited to classify data as being in the gas, liquid phase, or regions of the SC state. Considering clusters as

fluid regions with similar behavior, we provide symbolic expressions for the respective diffusion coefficients as function of density and temperature for each cluster and argue on their applicability and generalizability in physical terms.

In order to make an initial separation of the initial data to feed into the SR model, the process of hierarchical cluster analysis has been used for data classification into groups of similar objects/behavior[25,26]. More specifically, the agglomerative hierarchical clustering builds a cluster hierarchy displayed as dendrogram, and this is represented through a clustergram, i.e., a two-fold graphic representation of hierarchical clustering and heatmap, where colors are arranged according to the similarity of their measure. Hierarchical agglomerative clustering (bottom-up dendrogram) produces 1 to n clusters, where n is the total number of observations (in our case n=927). Cluster similarity increases as one goes down from 1 cluster (contains all the objects) to n clusters (each observation is its own cluster).

Another popular family of clustering comes from the *k*-type algorithms. The *k*-means incorporates k cluster means for data, assigns points to the nearest mean in terms of the Euclidean distance, and recalculates the mean of each cluster, until it reaches the desired number of clusters. The *k*-medoids scheme, on the other hand, incorporates k representative objects (the objects with the smallest sum of dissimilarities) from the input data (the medoids) as the calculation basis and is not limited to the Euclidean distance function, for example, it can also function with squared Euclidean or Manhattan distance[51]. Here we have also incorporated the k-medoids scheme and found similar results to agglomerative clustering.

D. Symbolic Regression

SR is an implementation of Genetic Programming[14], where a given problem is mapped to a sequence of parent-child tree structures, following Darwin's theory of evolution. Tree nodes are

mathematical operators and leaf nodes correspond either to input variables or constants[52]. An initial tree scheme (parent) is randomly considered at the beginning of the evolution process and various implementations of child structures follow, where a node or a branch of nodes is substituted in the next instance, either by preserving a parent's characteristic (cross-over) or by imposing totally new functions (mutation), as shown in Fig. 1b. Each proposed scheme is evaluated on accuracy (MSE) and complexity (the induced tree depth) measures and successful ones, i.e., those with low complexity and minimum error, lie at the Pareto front[53], and are stored as possible output expressions.

Complexity is closely connected to the number of nodes used in the SR tree and in this work spans over $1 \leq \mathrm{Comp.} \leq 20$. The available set of mathematical operators includes $\{ADD, SUB, MUL, DIV, POW\}$, functions $\{EXP, LOG, SIN, COS\}$, variables $\{\rho^*, T^*\}$ and any constant that may be needed. A further criterion used to select the output equation is the Akaike information criterion (AIC)[54]. AIC is well suited to predictive modeling applications and provides the means of estimating the unknown parameters of a given model based on the Maximum Likelihood Principle, by searching of the best candidate model in the sense of Kullback–Leibler information[55]. In most cases, the best model to choose from those suggested is the one with the smallest AIC value, positive or negative[56].

We have to note that the choice of simpler expressions, apart from computational gain and readable physical explanation[57,58], are less sensitive to overfitting[59] and open for interpretation[60]. Nevertheless, potential solutions to the problem have to be restricted by physical intuition apart from error and complexity metrics. The suggested expression must be evaluated in terms of physical correspondence to the problem, since the SR algorithm cannot distinguish the realizable model from a group of suggested functions. It is pointed out that the loop-nature of SR outputs a vast number of expressions, those with poor metrics are rejected, and human intervention is needed to keep only those that satisfy physical laws, at least when the model behavior is known or can be inferred.

Parallel calculations and increased computational load are inherent characteristics of the SR algorithm. Here we have implemented the Julia-language version by Cranmer et. al[20], which has given accurate and physical-based results in similar problems[35].

III. RESULTS

A. Unsupervised learning

The agglomerative clustering has been applied to the original dataset of $D^* = f(\rho^*, T^*)$, supposing that each data point is defined as a separate group/cluster. Data points have been gathered from relevant literature sources[61–77]. Through an iterative process, in every step, two similar clusters are spotted and joined together into a new cluster. The process continues until there is only one cluster containing the entire data set. The average linkage hierarchical clustering algorithm is applied for data classification.

In a dendrogram, the vertical axis represents the distance of similarity or dissimilarity between clusters, i.e., points that are joined lower on the vertical axis have more similarity that other that are joined at a higher value of the vertical axis. If two objects are most similar, the height of the link that joins them together is the smallest at the vertical axis. Overall, the height of the dendrogram indicates the order in which the clusters were joined. The number of the clusters is found using the elbow method[78], where in our dataset has discovered six data groups.

Based on the distribution of the measure values (using the average-linkage clustering) we see how groups are formed together. The hierarchy built by the clustering algorithm is represented by the clustergram given in Fig. 3a, where two dendrograms are presented. The horizontal dendrogram shows the clustering of the group based on the three variables-measures (D^*, ρ^*, T^*), and the vertical clustering indicates how these variables are joined together. At each dendrogram one axis refers to clustering of the object and the vertical axis refers to the distance similarity – dissimilarity. According

to the performance of the elbow criterion, where it is suggested to divide the data into six groups, in Fig. 3a, we observe (from left to right) six main clusters as six branches with different colors (purple, cyan, black, green, red, and blue). Clustergram colors depict how these variables contribute to the construction of the horizontal dendrogram, as well as to the creations of the vertical cluster.

Characteristics of the input dataset and the proposed cluster division are given in Table 1. In search of physical meaning to the imposed division, we note that the purple cluster (Fig. 3b) ~~group~~ entails the gas state and the SC gas-like region above it (see Fig. 2a), where temperature is higher than the critical value. Therefore, it is expected that the purple region might act as gas-like. Taking a closer look in Fig. 2a from left to right, we observe that the cyan and black regions overlap the line that extends above the critical-point of the LJ fluid[36]. These two clusters are prone to ambiguous behavior since they lie in the proximity of the critical-point, at the left-side of the Widom line, which demarcates the SC LJ space into liquid-like (on the right) and vapor-like regions (on the left)[49]. Another SC region division is suggested by the Frenkel line[79], which further divides the SC-liquid regions to “rigid” and “nonrigid” liquid, a fact that leads to qualitative change in many important fluid properties. The Frenkel line is drawn by approximation, in order to qualitatively depict the different regions that are possible to lie on its left and right side.

Most data come from the SC region, which is mainly represented by the green cluster. A small SC part belongs to the red cluster, which may be considered as liquid-like region, with presence of density fluctuations and/or formation of particle clusters that lead to local inhomogeneities[80,81]. Red points can be also found inside the pure liquid region, while the blue cluster should be seen as liquid/dense-liquid region. There are some sparse points in the gas-liquid co-existence region and the solid region, however, we do not discard them from the calculations. Conclusively, it is derived that the proposed clustering has satisfyingly captured the physical behavior of the LJ fluid. Overlapping regions, if not due to computational errors, are to be further examined by the clustergrams.

FIG. 2. Clusters formed by the agglomerative method. The six clusters are shown in different color and being compared to known fluid states. a) (ρ^*, T^*) phase diagram. From left to right we correlate with fluid states as: gas/SC-gas, SC-liquid non-rigid, SC-liquid rigid, liquid, dense liquid/solid, solid. Lines that separate fluid states and critical LJ values are taken from[36] and the Frenkel line is drawn by approximation with data from[79], b) (ρ^*, T^*, D^*) 3D plot, correlating each cluster diffusion coefficient values with LJ fluid state.

TABLE 1. Value range of the input dataset and the proposed cluster division. The number of samples for each cluster, *N*, is also given.

Cluster	N	ρ^*	T^*	D^*
Purple	97	0.005-0.075	0.700-6.004	2.014-31.808
Cyan	59	0.040-0.150	1.202-6.004	1.187-8.181
Black	111	0.100-0.300	1.250-6.003	0.574-4.749
Green	407	0.175-1.400	0.750-1.000	0.054-1.700
Red	195	0.600-1.171	0.750-2.500	0.013-0.182
Blue	62	0.750-1.050	0.530-1.000	0.001-0.069

In Fig. 2b we project (ρ^*, T^*) space to the respective diffusion coefficient values. D^* at gas state (i.e., purple and parts of the cyan and black clusters) comes with extra high values as expected and it is mainly affected by temperature values, as the density range is narrow at these regions. The green cluster extends almost over the whole phase space. Here, liquid-like behavior is prominent as density increases, and this is the reason for obtaining smaller diffusion values. The red (liquid) and blue

(liquid/solid) clusters are examined in a very narrow temperature range, and present small diffusion coefficient values.

As far as the horizontal dendrogram is concerned, Fig. 3a shows the hierarchical clustering of all data points, while Figs. 3b-g present the six distinct clusters (groups), drawn in different colors for visibility reasons (light blue, red, green, black, cyan and purple). Clustergram colors depict how these variables contribute to the construction of the horizontal dendrogram, as well as to construction rules for the vertical cluster. For example, in Figs. 3b, f, and g, the diffusion coefficient D^* has the main contribution to the formation of each group (purple, red and blue, respectively) compared to the other two parameters (ρ^*, T^*). Also, T^* is the dominant criterion to the formation of the black and green cluster, while ρ^* is responsible for the formation of the cyan cluster.

FIG. 3. Results for unsupervised learning**,** a) agglomerate clustering output, in six clusters, and partial investigation on each one, b) purple cluster (gas/SC gas), c) cyan (SC gas) and d) black cluster (SC gas/SC non-rigid liquid), e) green cluster (SC rigid/non-rigid liquid), red cluster (SC rigid liquid), and f) blue cluster (liquid/dense liquid).

B. Symbolic equations for each cluster

Literature results have shown that Stokes-Einstein-Sutherland[82–84] (SES) empirical equations have shown good fit on diffusion data[85]. In this work, we have performed a partial investigation, where each cluster is treated as an independent fluid region, with different properties, and a symbolic expression is extracted for each one.

The purple cluster corresponds to gas diffusion coefficient values in the range $D^* = \{2.01 - 31.81\}$. To capture diffusion behavior in such a wide value range, our method has proposed the equations shown in Fig. 4, both of which reveal that diffusion is inversely depended on fluid density, ρ^*, and linearly depended on T. Similar behavior has been presented in a previous work for the LJ gas-only state[35], while, for real fluids, in a mixture of ibuprofen and liquid ethanol in supercritical CO_2 , D values have been also found to decrease with increasing density and increased with increasing temperature[46]. Metrics for each equation are tabulated in Table 2. Accuracy, complexity, and AIC metrics are close for both equations.

The cyan cluster corresponds to gas diffusion coefficient values in the range $D^* = \{1.19 - 8.18\}$. Here there is still strong effect of temperature (Fig. 5) and inverse density effect. Especially for the second equation proposed, $D^*_{c_2} = T^* - A\rho^* + B$, which has given a smaller AIC value, ρ^* and T^* are linearly affecting D^*, though in contradictory manner. Notwithstanding the fact that this is a straightforward effect, the neighboring black cluster presents more complex behavior (Fig. 6), as approached by the equation $D^*_{k_2} = (AT^* + B)(C - D\rho^*)$, which represents the increase of D^* with the increase of T^*, the decrease of D^* with increasing ρ^*, along with a combined reduction of D^* due to T^* and ρ^*. We have to point out that the black cluster overlaps the Widom line and extends above the critical point. It has been argued that this critical region is associated with large density fluctuations and affects measured diffusion coefficients both in simulations and experiments[85]. Simulation results are characterized by statistical uncertainties and experiments by larger errors.

FIG. 4. Identity plot for the two equations extracted by the SR method for the purple cluster. SR-derived data D^*_{pred} from the proposed equations a) $D^*_{p_1} = A \cdot \frac{T^*+B}{\rho^*}$ a and b) $D^*_{p_2} = \frac{A \cdot T^*}{\rho^*(T^*+B)}$ (see Table 2) are compared to database data D^*_{calc} on the 45° dotted line.

FIG. 5. Identity plot for the two equations extracted by the SR method for the cyan cluster. SR-derived data D^*_{pred} from the proposed equations a) $D^*_{c_1} = \frac{T^*}{\rho^* + A}$ and b) $D^*_{c_2} = T^* - A\rho^* + B$ (see Table 2) are compared to database data D^*_{calc} on the 45° dotted line.

The green region covers the largest percentage of the dataset, on the right side of the Frenkel line, and lie in the SC-liquid state, giving diffusion coefficient values in a small range, as $D^* = \{0.0054 - 1.7\}$. Figure 7 presents the two proposed equations. These are both of higher complexity (Comp.) compared to the more gas-like cluster equations shown previously. We attribute this to the fact that there are sub-regions within the green cluster that affect D^* in contradicting ways, as also depicted by the partial clustergram in Fig. 3e. The self-diffusion coefficient vs. density at constant temperatures has shown an inflection near the Frenkel line[48]. The equations for green region have to encapsulate all effects in one expression.

FIG. 6. Identity plot for the two equations extracted by the SR method for the black cluster. SR-derived data D^*_{pred} from the proposed equations a) $D^*_{k_1} = T^*(A - \rho^*)$ and b) $D^*_{k_2} = (AT^* + B)(C - D\rho^*)$ (see Table 2) are compared to database data D^*_{calc} on the 45° dotted line.

FIG. 7. Identity plot for the two equations extracted by the SR method for the green cluster. SR-derived data D^*_{pred} from the proposed equations a) $D^*_{g_1} = A\frac{T^*}{\rho^*} - B\rho^* + C$ and b) $D^*_{g_2} = A\frac{T^* + B}{(\rho^*)^2 - C\rho^* + D}$ (see Table 2) are compared to database data D^*_{calc} on the 45° dotted line.

For the red cluster, the decrease of diffusion coefficients is prominent, and attributed to the increasing fluid density in liquid state, which leads to increased particle collision[44]. Temperature effects have been diminished here, and this is well-discovered by the first SR equation, $D^*_{r_1} = e^{-A\rho^*}$, in which there is only dependence on density, in an exponential manner. However, to increase accuracy, one has to take into account the small dependence of temperature, as suggested by the second equation, $D^*_{r_2} = e^{AT^*-B\rho^*}$ (Fig. 8). Another issue noticed here is that diffusion values in this data cluster, as is the case for all pure liquids, are given in a small temperature range and this can be the reason that the equations have extracted only small effect of T^* on D^*.

FIG. 8. Identity plot for the two equations extracted by the SR method for the red cluster. SR-derived data D^*_{pred} from the proposed equations a) $D^*_{r_1} = e^{-A\rho^*}$ and b) $D^*_{r_2} = e^{AT^*-B\rho^*}$ (see Table 2) are compared to database data D^*_{calc} on the 45° dotted line.

At even denser liquids, with solidification effects, a change of T^* on D^* at constant ρ^* is even smaller (blue region)[45]. Literature results have given this dependence of diffusion to temperature analog to $\sqrt{T^*}$ [47]. The respective SR equations in Fig. 9 have given MSE values practically zero, as differences in $D^* = \{0.0006 - 0.069\}$ refer to very small values.

FIG. 9. Identity plot for the two equations extracted by the SR method for the blue cluster. SR-derived data D^*_{pred} from the proposed equations a) $D^*_{b_1} = AT^*(B - \rho^*)$ and b) $D^*_{b_2} = -\frac{AT^*}{\rho^*}(B\rho^* - C)$ (see Table 2) are compared to database data D^*_{calc} on the 45° dotted line.

TABLE 2. Cluster-specific, SR-extracted formulas for the self-diffusion coefficient, with metrics of complexity (Comp.), accuracy (R^2), error estimation (MSE, MAE), and the Akaike criterion to select the best-fit equation (AIC).

cluster	Equation	Comp.	R^2	MAE	MSE	AIC
Purple	$D^*_{p_1} = A \cdot \frac{T^* + B}{\rho^*}$ $A = 0.103104$ $B = 0.353675$	7	0.99	0.4994	0.5913	224.39
	$D^*_{p_2} = \frac{A \cdot T^*}{\rho^*(T^* + B)}$ $A = 1.79437$ $B = 11.684$	9	0.99	0.3435	0.5249	214.35
Cyan	$D^*_{c_1} = \frac{T^*}{\rho^* + A}$ $A = 0.755629$	5	0.95	0.3385	0.1918	71.41
	$D^*_{c_2} = T^* - A\rho^* + B$ $A = 14.1172$ $B = 1.90849$	7	0.97	0.2191	0.1312	49.56
Black	$D^*_{k_1} = T^*(A - \rho^*)$ $A = 0.698033$	5	0.90	0.2426	0.1105	63.03

	$D^*_{k_2} = (AT^* + B)(C - D\rho^*)$ $A = 0.728116$ $B = 0.531705$ $C = 1, D = 2$	11	0.93	0.1531	0.0736	27.39
Green	$D^*_{g_1} = A\frac{T^*}{\rho^*} - B\rho^* + C$ $A = 0.049446$ $B = 0.530784$ $C = 0.469847$	11	0.95	0.0483	0.0073	-844.77
	$D^*_{g_2} = A\frac{T^* + B}{(\rho^*)^2 - C\rho^* + D}$ $A = 0.015653$ $B = 1.219230$ $C = 0.568809$ $D = 0.159102$	15	0.97	0.0376	0.0045	-1046.29
Red	$D^*_{r_1} = e^{-A\rho^*}$ $A = 3.108978$	4	0.91	0.0133	0.0003	-1056.43
	$D^*_{r_2} = e^{AT^* - B\rho^*}$ $A = 3.108978$ $B = 4.391156$	8	0.98	0.0071	0.0001	-1283.28
Blue	$D^*_{b_1} = AT^*(B - \rho^*)$ $A = 0.371697$ $B = 0.984116$	7	0.93	0.0053	<0.00001	-453.27
	$D^*_{b_2} = -\frac{AT^*}{\rho^*}(B\rho^* - C)$ $A = 0.230995$ $B = 1.04979$ $C = 1.05966$	11	0.95	0.0036	<0.00001	-481.26

C. Comparison with literature results

The proposed SR equations from this work are compared to well-established empirical relations from the literature. We have to point out that Eqs. (1-3) have been extracted in order to describe specific regions of the (ρ^*-T^*) space. More specifically, Eq. 1 refers to a HS model and has been proposed by Chapman and Cowling[39] for dilute gases. Equation 2 is a revised form of Eq. 1 and has been given by Speedy[40], though still based on HS model. Zhu et. al[30] have proposed Eq. 3, which is in turn an improvement of Eq. 2.

Figure 10 presents cluster-specific fits of Eqs. (1-3) and the respective SR relations from Table 2. The diffusion coefficient values are plotted vs. the number of samples in the dataset (N). For the purple cluster, in Fig. 10a, which encompasses gas and SC gas states, $D^*_{p_1}$ fits perfectly to the original dataset, while all empirical Eqs. (1-3) follow the data trend but overestimate the values of diffusion coefficient. The critical data range near the Widom line (cyan and black cluster, Figs. 10b-c) fits well on Eqs. 2-3 and the proposed SR equations. Equation 1, as it has been proposed for gas-like fluids, fails to keep pace with data from all remaining clusters and it presents a bias that sets it above data values. Data corresponding to various liquid-like states of the green and the red cluster are approached by Eqs. 2-3, but only $D^*_{g_2}$ and $D^*_{r_2}$ fit well. The dense-liquid, blue cluster is fitted by $D^*_{b_2}$ and, to some extent, by Eq. 2.

FIG. 10. Comparison to empirical Eq. (1)[39], Eq. (2)[40], and Eq. (3)[30] vs. the number of samples, N, in each cluster with a) $D^*_{p_1}$ for the purple cluster (gas/SC gas), b) $D^*_{c_2}$ for the cyan cluster (SC gas), c) $D^*_{k_2}$ for the black cluster (SC gas/SC non-rigid liquid), d) $D^*_{g_2}$ for the green cluster (SC rigid/non-rigid liquid) , e) $D^*_{r_2}$ for the red cluster (SC rigid liquid), and f) $D^*_{b_2}$ for the blue cluster (liquid/dense liquid). All data is sorted for visibility reasons.

IV. DISCUSSION

Starting from a diffusion coefficients dataset, taken from relevant literature references, this work has focused on extracting analytical equations for the LJ fluid diffusion coefficients, keeping in mind that a proposed equation should be simple, accurate, and bound on firm physical laws. In a relevant work, it has been shown that self-diffusion coefficients of the Lennard-Jones (LJ) fluid can be derived as function of density and temperature for all fluid states[35]. In terms of more detailed approach, investigation between gas, liquid, and SC states (the region division has been based on literature information[36]) has also given region-specific expressions.

On a different view, if data are seen as the only source of information to take into consideration, unsupervised learning techniques can be deployed to define fluid regions of similar behavior. The novel contribution of this work is that the incorporation of unsupervised learning, with aid of the agglomerative hierarchical clustering, has revealed important features of the LJ fluid and, more interestingly, it has proposed data division in regions of similar behavior that resemble our fundamental physics understanding of treating the fluid state space in gas, liquid, and SC states. Moreover, intermediate regions of gas/liquid/solid co-existence under extreme temperature and/or density conditions have been spotted, and lead to further fluid division based on its position on the state space, as it is artificially demarcated by the Widom or the Frenkel line.

Passing all this information to our SR methodology, it has been found that the diffusion coefficient presents similar behavior in gas and SC-gas only states. Although being in extreme conditions, the effect of temperature is prominent, and, as it spans over $T^* = \{1.0 - 6.0\}$, diffusion coefficients lie in the range $D^* = \{2.01 - 31.81\}$. From the two symbolic expressions extracted by the SR process presented in Table 2, the first choice would be the one with the simplest form, $D^*_{p_1}$, while $D^*_{p_2}$ would be better incorporated in applications where accuracy is the question.

While arguing on the critical fluid region above and near the Widom line, our method has verified the existence of a region where large density variations exist, along with statistical

uncertainties and larger errors[85]. Moreover, the "rough" estimation of the Frenkel line has been presented as the limit between two consecutive clusters, in which the LJ fluid presents different behavior. As such, the SC, liquid-like fluid is further divided in rigid/non-rigid regions. These two states correspond to the point that the liquid leaves its harmonic, solid-like motion of small diffusion and enters the gas-like, large-amplitude, ballistic-collisional motion[79].

The hierarchical clustering method has inferred that the cyan and black clusters are prone to significant property variations and are mainly affected by temperature. Gas-like behavior, though with liquid co-existence, has been observed in these clusters, and as expected, the diffusion coefficients attain high values in the range $D^* = \{0.57 - 8.18\}$. The Frenkel line poses a new boundary of stronger "liquid-like" behavior that refers to the green cluster. However, the green cluster, which occupies most of the available state space, presents subregions of different behavior. At liquid-like regions, the diffusion coefficients are decreasing along with particle mobility, and collision happen more often. There is small contribution from temperature and density effects have the defining role.

Empirical models, such as the smooth/rough HS models or the SES equation, have failed to reproduce experimental measurements at high-pressures for molecular liquids, such as methane[86]. In a more recent work, it is also suggested that, while the SES relation holds in a wide pressure and density regime and may be incorporated in cases where experimental data are missing, it fails at highest densities in the vicinity of the freezing point[87]. In this work we have not investigated the application of SES to the dataset, as it is in the form of $D \propto 1/\eta$, and shear viscosity, η, data are missing. Nevertheless, widely used empirical diffusion coefficient models, with dependance on T^* and ρ^*, have been investigated and compared to the proposed equations extracted from SR methods.

It has become clear that the proposed, cluster-specific SR equations proposed (see Table 2) have shown excellence performance by capturing LJ fluid data behavior across the whole phase space, and, if their low complexity is also taken into account, they seem to be a valuable alternative to

describe mass transport phenomena through the diffusion coefficient. In contrast, the empirical Eqs. (1-3), have been accurate only for specific parts of the dataset and cannot be considered to describe fluid properties across all fluid states, in ambient or extreme conditions. This highlights the lack of a universal prediction model, and the proposed SR-extracted partial equations are posed as valuable alternatives that fit almost perfectly on provided data.

The method presented in this work employs data referring to the LJ fluid. It would be of interest to extrapolate our findings to real fluids and mixtures. Usually, this can be done with the incorporation of the LJ parameters (i.e., ε and σ) inside the symbolic expressions and the calculation of new critical values for temperature and density[30]. Efforts towards this direction have been made, but more simulation and experimental data are needed before we reach in safe conclusion on the final form of the equations.

V. CONCLUSIONS

We have introduced a machine learning and statistical framework and managed to dive deep into physics concepts without any prior knowledge of the system under investigation. Symbolic Regression has been investigated on its ability to discover distinct regions of the search space by superimposing neighborhood criteria obtained via clustering. Notably, our unsupervised procedure is able to recognize trends and different behaviors from experimental and simulation datasets. More interestingly, it has been shown that SR procedures can elevate computational science to a new level, where analytical equations, with practical and physical meaning, are at hand, performing better than well-established empirical equations. The combination with clustering techniques may reveal the underlying differences in the system described in various regions and, thus, employ specific equations to describe and predict their properties. The ML “black” box is uncovered, and these approaches can have a leading role in predicting and guiding computational science. Further

investigation on real fluids and mixtures and generalization of such methods to neighboring fields, computationally demanding and difficult to interpret, are the future objectives.

CONFLICTS OF INTEREST

The authors have nothing to disclose.

AUTHOR'S CONTRIBUTIONS

F.S. conceptualization, writing, programming, editing. K.P. clustering, programming, writing. A.C. clustering, programming, writing. T.E.K. writing, review and editing.

ACKNOWLEDGEMENT

F.S. acknowledges support by the Center of Research Innovation and Excellence of University of Thessaly, funded by the Special Account for Research Grants of University of Thessaly (project CAMINOS, No. 5600.03.08.03). We also acknowledge computational time granted from the National Infrastructures for Research and Technology S.A. (GRNET S.A.) in the National HPC facility - ARIS.

DATA AVAILABILTY

The dataset incorporated in this study has been downloaded from https://aip.scitation.org/doi/suppl/10.1063/5.0011512.

References

[1] F. Sofos, C. Stavrogiannis, K.K. Exarchou-Kouveli, D. Akabua, G. Charilas, and T.E. Karakasidis, Fluids **7**, 116 (2022).

[2] J. Wei, X. Chu, X. Sun, K. Xu, H. Deng, J. Chen, Z. Wei, and M. Lei, InfoMat **1**, 338 (2019).

[3] N.H. Paulson, S. Zomorodpoosh, I. Roslyakova, and M. Stan, Calphad **68**, 101728 (2020).

[4] G.C.Y. Peng, M. Alber, A. Buganza Tepole, W.R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W.W. Lytton, P. Perdikaris, L. Petzold, and E. Kuhl, Arch. Comput. Methods Eng. **28**, 1017 (2021).

[5] D. Stephenson, J.R. Kermode, and D.A. Lockerby, Microfluid. Nanofluidics **22**, 1 (2018).

[6] A.P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Phys. Rev. X **8**, 041048 (2018).

[7] B. Mortazavi, E. V. Podryabinkin, S. Roche, T. Rabczuk, X. Zhuang, and A. V. Shapeev, Mater. Horiz. **7**, 2359 (2020).

[8] W. Sun, Y. Zheng, K. Yang, Q. Zhang, A.A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, and K. Sun, *Machine Learning-Assisted Molecular Design and Efficiency Prediction for High-Performance Organic Photovoltaic Materials* (2019).

[9] F. Sahli Costabal, K. Matsuno, J. Yao, P. Perdikaris, and E. Kuhl, Comput. Methods Appl. Mech. Eng. **348**, 313 (2019).

[10] G.T. Craven, N. Lubbers, K. Barros, and S. Tretiak, J. Chem. Phys. **153**, (2020).

[11] A. Kashefi, D. Rempe, and L.J. Guibas, Phys. Fluids **33**, 027104 (2021).

[12] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, and H. Pérez-Sánchez, Drug Discov. Today **23**, 1784 (2018).

[13] S. Bhattacharya, M.K. Verma, and A. Bhattacharya, Phys. Fluids **34**, 025102 (2022).

[14] J.R. Koza, Stat. Comput. 1994 42 **4**, 87 (1994).

[15] S.-M. Udrescu and M. Tegmark, Sci. Adv. **6**, eaay2631 (2020).

[16] C. Loftis, K. Yuan, Y. Zhao, M. Hu, and J. Hu, J. Phys. Chem. A **125**, 435 (2021).

[17] J.P.S. Aniceto, B. Zêzere, and C.M. Silva, J. Mol. Liq. **326**, 115281 (2021).

[18] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, and L. Ward, Mol. Syst. Des. Eng. **3**, 819 (2018).

[19] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, and L.M. Ghiringhelli, Phys. Rev. Mater. **2**, 083802 (2018).

[20] M. Cranmer, A. Sanchez Gonzalez, P. Battaglia, R. Xu, K. Cranmer, D. Spergel, and S. Ho, in *Adv. Neural Inf. Process. Syst.*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Curran Associates, Inc., 2020), pp. 17429–17442.

[21] S.L. Brunton, B.R. Noack, and P. Koumoutsakos, Annu. Rev. Fluid Mech. **52**, 477 (2020).

[22] Y. Long, J. Ren, and H. Chen, Phys. Rev. Lett. **124**, 185501 (2020).

[23] O. Balabanov and M. Granath, Phys. Rev. Res. **2**, 013354 (2020).

[24] J.F. Rodriguez-Nieva and M.S. Scheurer, Nat. Phys. **15**, 790 (2019).

[25] F. Murtagh and P. Contreras, WIREs Data Min. Knowl. Discov. **2**, 86 (2012).

[26] F. Murtagh and P. Contreras, ArXiv11050121 Cs Math Stat (2011).

[27] George. Karniadakis, A. Beşkök, and N.Rao. Aluru, *Microflows and Nanoflows : Fundamentals and Simulation* (Springer, 2005).

[28] F. Sofos, T.E. Karakasidis, and A. Liakopoulos, Nanosci. Nanotechnol. Lett. **5**, 457 (2013).

[29] A. Liakopoulos, F. Sofos, and T.E. Karakasidis, Phys. Fluids **29**, (2017).

[30] Y. Zhu, X. Lu, J. Zhou, Y. Wang, and J. Shi, Fluid Phase Equilibria **194–197**, 1141 (2002).

[31] A. Moradzadeh and N.R. Aluru, J. Phys. Chem. Lett. **10**, 1242 (2019).

[32] H.F. Ye, J. Wang, Y.G. Zheng, H.W. Zhang, and Z. Chen, Phys. Chem. Chem. Phys. **23**, 10164 (2021).

[33] F. Sofos and T.E. Karakasidis, Sci. Rep. **11**, (2021).

[34] J.P. Allers, J.A. Harvey, F.H. Garzon, and T.M. Alam, J. Chem. Phys. **153**, (2020).

[35] K. Papastamatiou, F. Sofos, and T.E. Karakasidis, AIP Adv. **12**, 025004 (2022).

[36] S. Stephan, M. Thol, J. Vrabec, and H. Hasse, J. Chem. Inf. Model. 4248 (2019).

[37] C. Zeni, K. Rossi, T. Pavloudis, J. Kioseoglou, S. de Gironcoli, R.E. Palmer, and F. Baletto, Nat. Commun. **12**, 6056 (2021).

[38] F. Casadei and G.L. Pappa, Nat. Comput. **20**, 753 (2021).

[39] S. Chapman, T.G. Cowling, D. Burnett, and C. Cercignani, *The Mathematical Theory of Non-Uniform Gases: An Account of the Kinetic Theory of Viscosity, Thermal Conduction and Diffusion in Gases* (Cambridge University Press, 1990).

[40] R.J. Speedy, Mol. Phys. **62**, 509 (1987).

[41] E.A. Ploetz and P.E. Smith, J. Phys. Chem. B **123**, 6554 (2019).

[42] V.V. Brazhkin, Yu.D. Fomin, A.G. Lyapin, V.N. Ryzhov, and E.N. Tsiok, J. Phys. Chem. B **115**, 14112 (2011).

[43] L.L. Williams, J.B. Rubin, and H.W. Edwards, Ind. Eng. Chem. Res. **43**, 4967 (2004).

[44] J.J. Suárez, I. Medina, and J.L. Bueno, Fluid Phase Equilibria **153**, 167 (1998).

[45] K.K. Liong, P.A. Wells, and N.R. Foster, J. Supercrit. Fluids **4**, 91 (1991).

[46] C.Y. Kong, K. Sugiura, S. Natsume, J. Sakabe, T. Funazukuri, K. Miyake, I. Okajima, S. Badhulika, and T. Sako, J. Supercrit. Fluids **159**, 104776 (2020).

[47] J.H. Dymond, J. Chem. Phys. **60**, 969 (1974).

[48] B. Scheiner and T.J. Yoon, J. Chem. Phys. **154**, 134101 (2021).

[49] F. Rahmani, T. Weathers, A. Hosangadi, and Y.C. Chiew, Chem. Eng. Sci. **214**, 115424 (2020).

[50] J. Peng, W. Wang, Y. Yu, H. Gu, and X. Huang, Chin. J. Chem. Phys. **31**, 404 (2018).

[51] E. Schubert and P.J. Rousseeuw, Inf. Syst. **101**, 101804 (2021).

[52] Y. Wang, N. Wagner, and J.M. Rondinelli, MRS Commun. **9**, 793 (2019).

[53] I. Giagkiozis and P.J. Fleming, Evol. Comput. **22**, 651 (2014).

[54] H. Akaike, IEEE Trans. Autom. Control **19**, 716 (1974).

[55] J.E. Cavanaugh and A.A. Neath, WIREs Comput. Stat. **11**, e1460 (2019).

[56] T.S. Baguley, *Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences.* (Palgrave Macmillan, New York, NY, 2012), pp. xxiii, 830.

[57] P.A.K. Reinbold, L.M. Kageorge, M.F. Schatz, and R.O. Grigoriev, Nat. Commun. **12**, 3219 (2021).

[58] V. Tavanashad, A. Passalacqua, and S. Subramaniam, Int. J. Multiph. Flow **135**, 103533 (2021).

[59] C. Loftis, K. Yuan, Y. Zhao, M. Hu, and J. Hu, J. Phys. Chem. A **125**, 435 (2021).

[60] G. Kronberger, L. Kammerer, B. Burlacu, S.M. Winkler, M. Kommenda, and M. Affenzeller, in *Genet. Program. Theory Pract. XVI*, edited by W. Banzhaf, L. Spector, and L. Sheneman (Springer International Publishing, Cham, 2019), pp. 85–102.

[61] see https://mmlapps.nist.gov/srs/LJ_PURE/md.htm for NIST, Molecular Dynamics Results., (n.d.).

[62] J. Kushick and B.J. Berne, J. Chem. Phys. **64**, 1362 (1976).

[63] P. Schofield, Comput. Phys. Commun. **5**, 17 (1973).

[64] J.P.J. Michels and N.J. Trappeniers, Chem. Phys. Lett. **33**, 195 (1975).

[65] J.P.J. Michels and N.J. Trappeniers, Phys. Stat. Mech. Its Appl. **90**, 179 (1978).

[66] S.-H. Chen and A. Rahman, Mol. Phys. **34**, 1247 (1977).

[67] K. Lucas and B. Moser, Mol. Phys. **37**, 1849 (1979).

[68] D.M. Heyes, J Chem Soc Faraday Trans 2 **79**, 1741 (1983).

[69] D.M. Heyes, Phys. Rev. B **37**, 5677 (1988).

[70] D.M. Heyes and J.G. Powles, Mol. Phys. **71**, 781 (1990).

[71] D.M. Heyes, J.G. Powles, and J.C. Gil Montero, Mol. Phys. **78**, 229 (1993).

[72] P. Borgelt, C. Hoheisel, and G. Stell, Phys. Rev. A **42**, 789 (1990).

[73] J.E. Straub, Mol. Phys. **76**, 373 (1992).

[74] R.L. Rowley and M.M. Painter, Int. J. Thermophys. **18**, 1109 (1997).

[75] M. Canales and J.A. Padró, Phys. Rev. E **60**, 551 (1999).

[76] K. Meier, A. Laesecke, and S. Kabelac, J. Chem. Phys. **121**, 9526 (2004).

[77] L. Wei-Zhong, C. Cong, and Y. Jian, Heat Transfer—Asian Res. **37**, 86 (2008).

[78] M.A. Syakur, B.K. Khotimah, E.M.S. Rochman, and B.D. Satoto, IOP Conf. Ser. Mater. Sci. Eng. **336**, 012017 (2018).

[79] V.V. Brazhkin, Yu.D. Fomin, A.G. Lyapin, V.N. Ryzhov, and K. Trachenko, Phys. Rev. E **85**, 031203 (2012).

[80] I. Skarmoutsos, D. Dellis, and J. Samios, J. Phys. Chem. B **113**, 2783 (2009).

[81] A. Idrissi, I. Vyalov, N. Georgi, and M. Kiselev, J. Phys. Chem. B **117**, 12184 (2013).

[82] W. Sutherland, Lond. Edinb. Dublin Philos. Mag. J. Sci. **9**, 781 (1905).

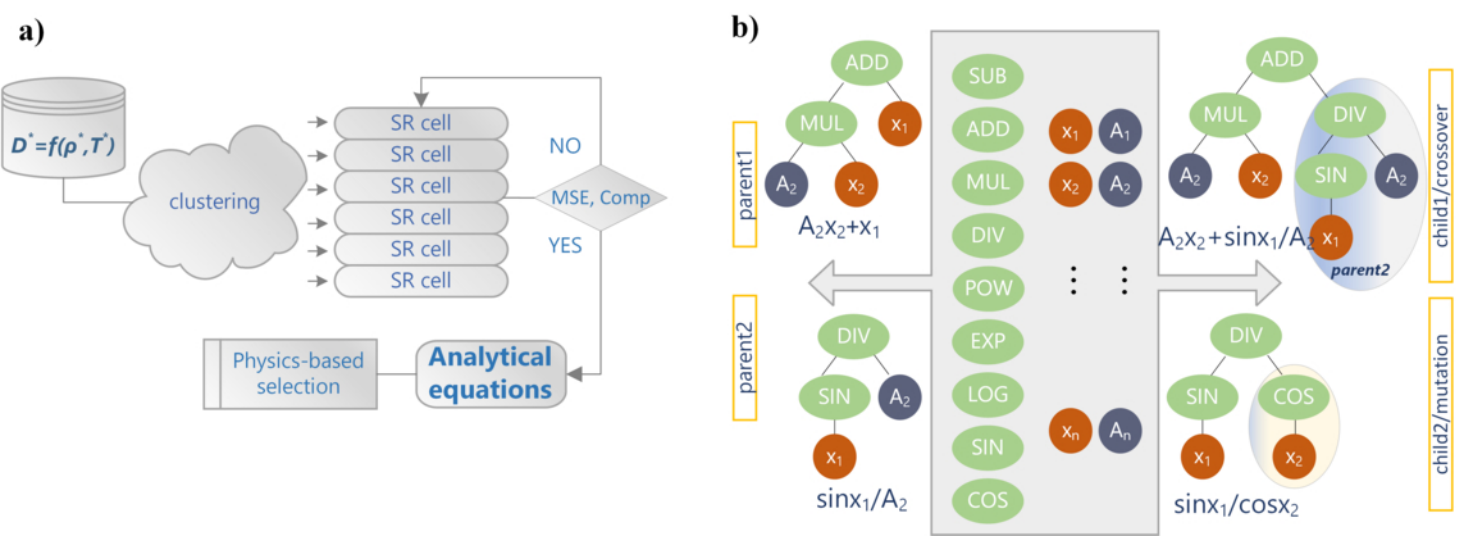
[83] A.E. Giannakopoulos, F. Sofos, T.E. Karakasidis, and A. Liakopoulos, Int. J. Heat Mass Transf. **55**, 5087 (2012).

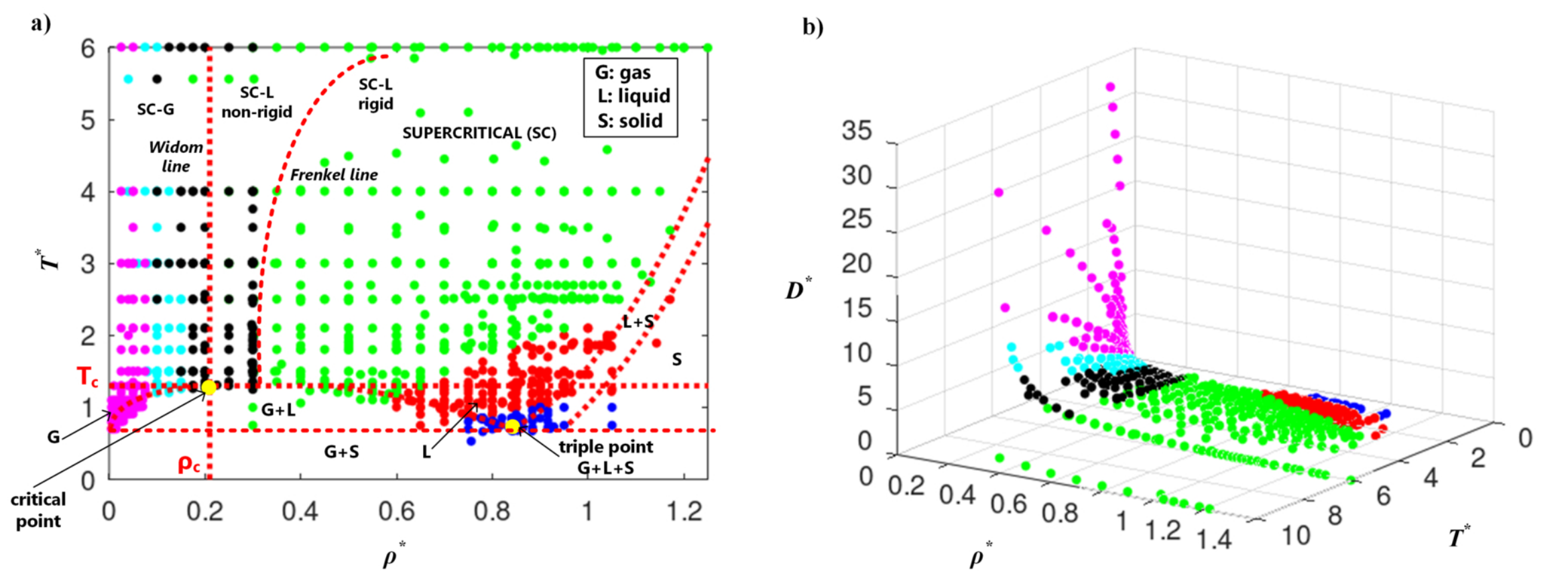
[84] R.H. Stokes, J. Am. Chem. Soc. **72**, 2243 (1950).

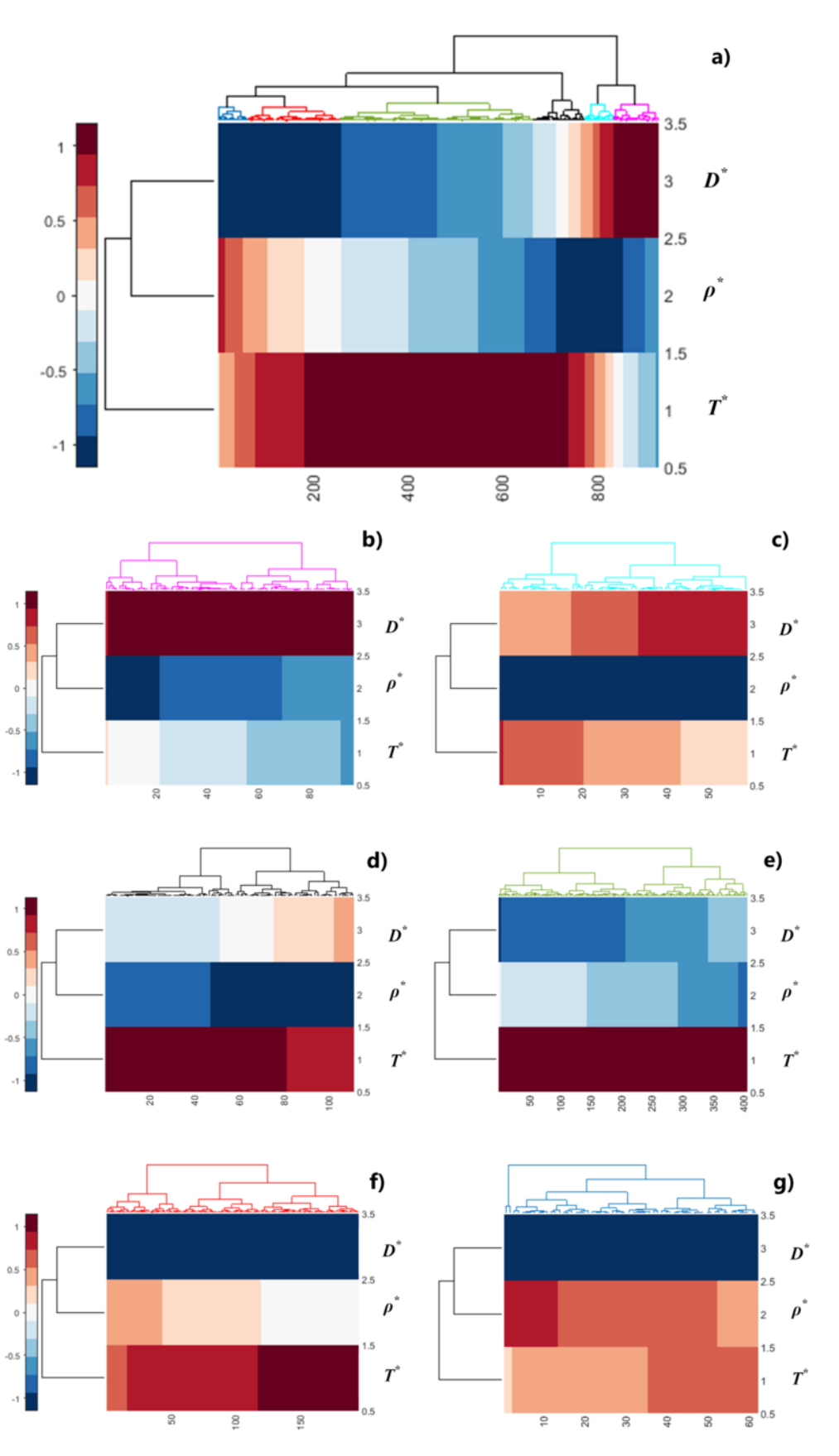
[85] G. Guevara-Carrion, S. Ancherbak, A. Mialdun, J. Vrabec, and V. Shevtsova, Sci. Rep. **9**, 8466 (2019).

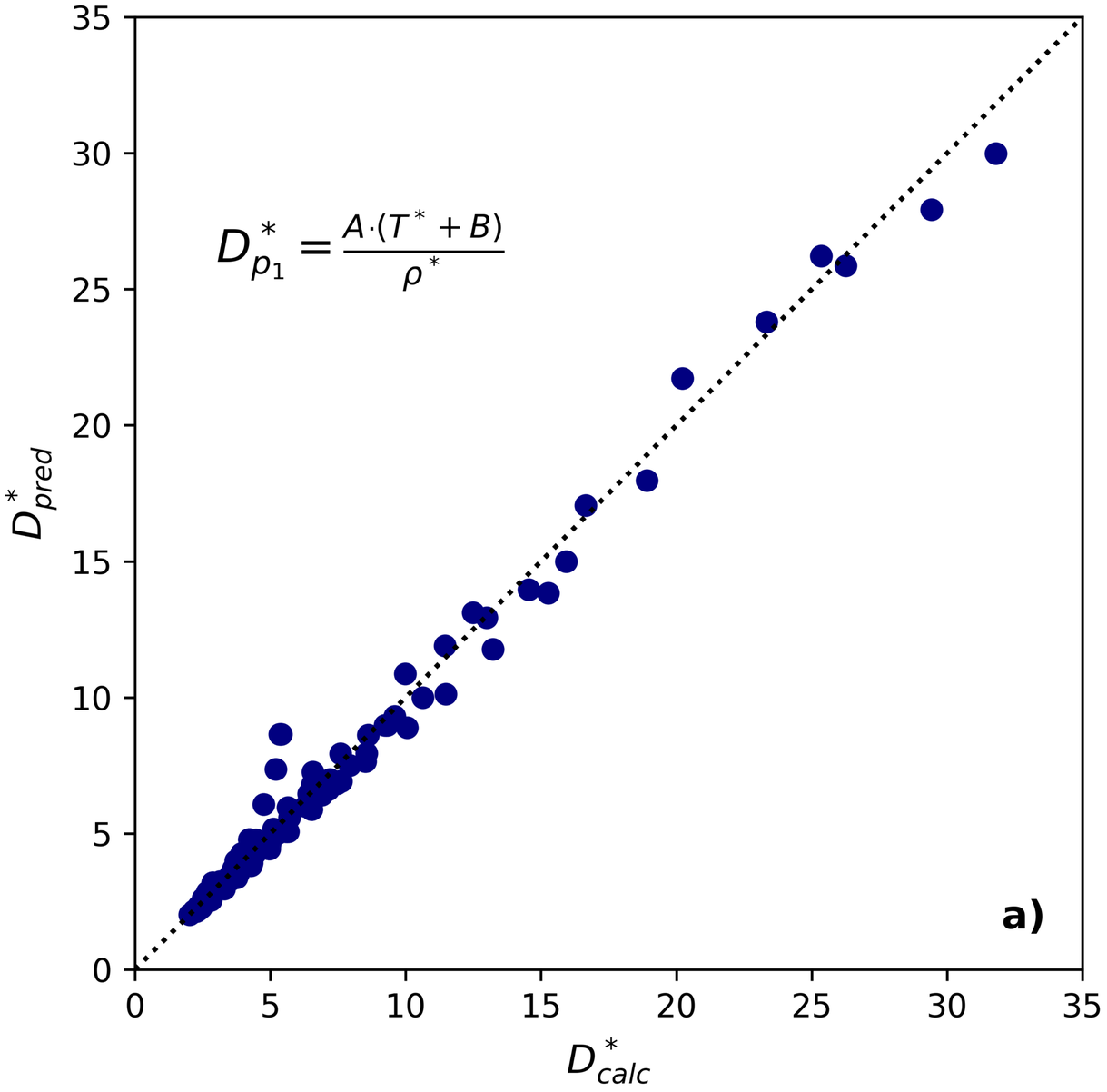
[86] U. Ranieri, S. Klotz, R. Gaal, M.M. Koza, and L.E. Bove, Nat. Commun. **12**, 1958 (2021).

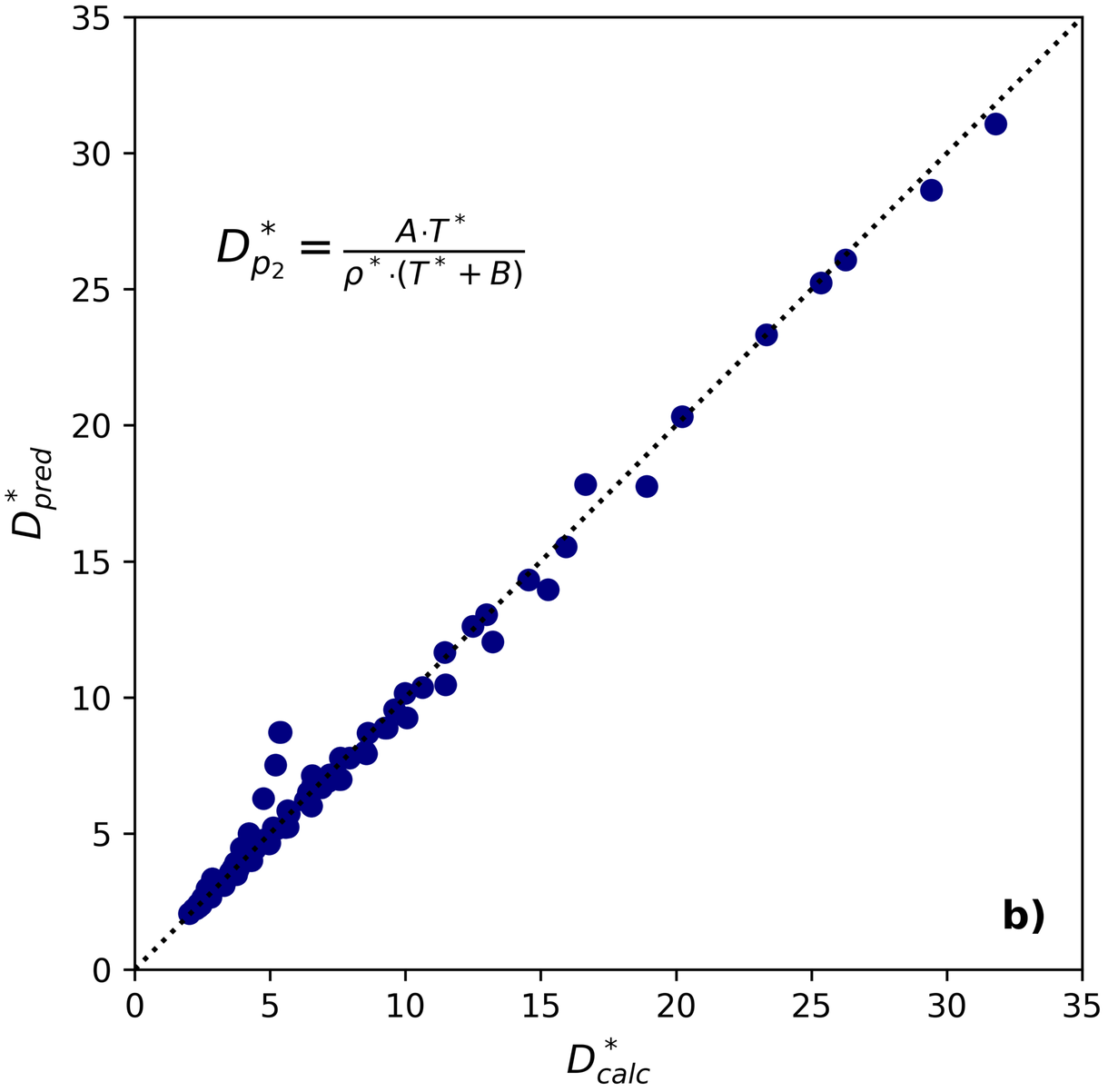
[87] S.A. Khrapak, J. Mol. Liq. **354**, 118840 (2022).

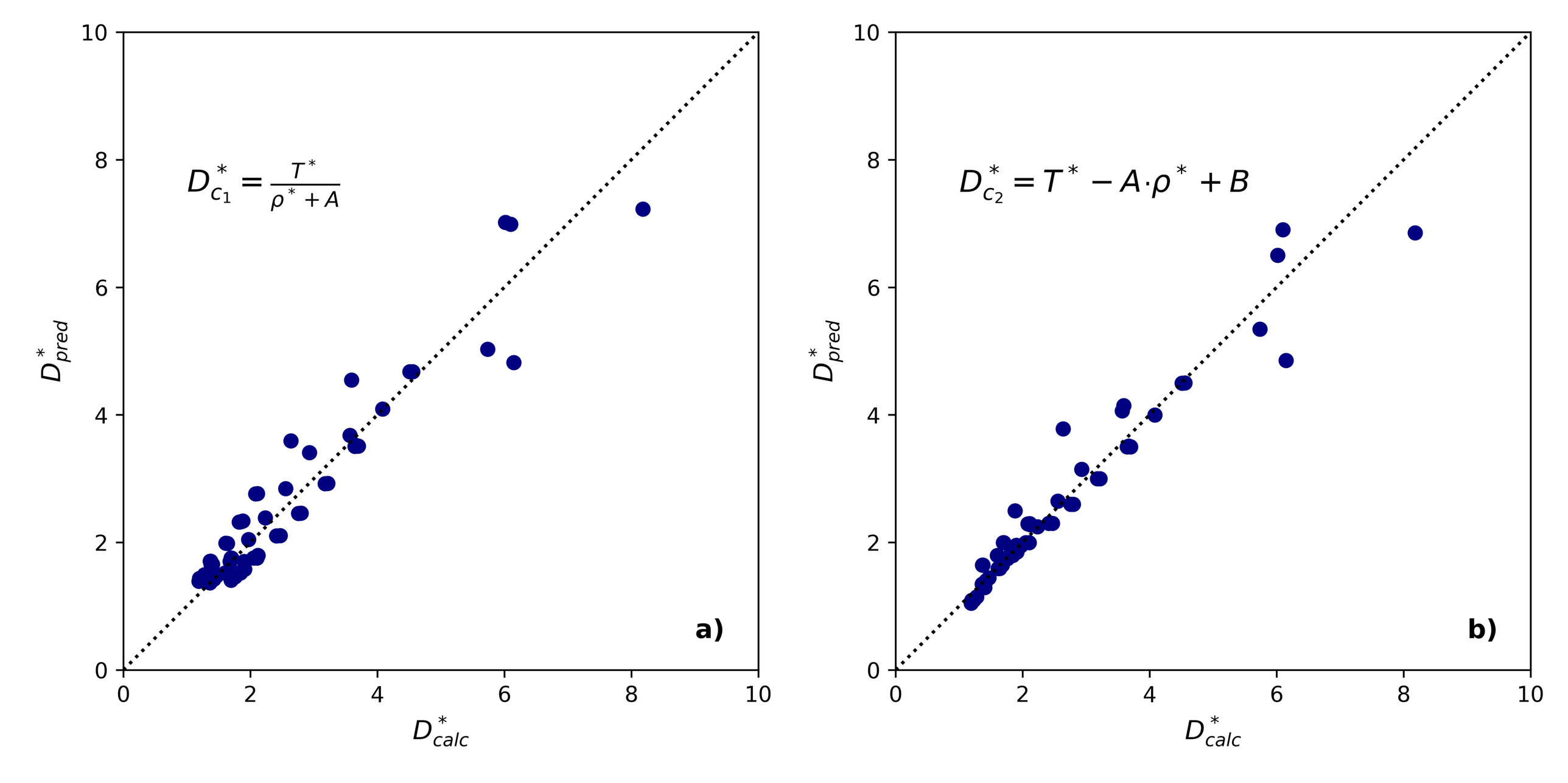

a)
D*=f(ρ*,T*)
clustering
SR cell
SR cell
SR cell
SR cell
SR cell
SR cell
NO
MSE, Comp
YES
Physics-based selection
Analytical equations
b)
parent1
parent2
ADD
MUL
x1
A2
x2
A2x2+x1
SUB
ADD
MUL
DIV
POW
EXP
LOG
SIN
COS
x1 A1
x2 A2
xn An
DIV
SIN
A2
x1
sinx1/A2
A2x2+sinx1/A2
child1/crossover
parent2
DIV
SIN
COS
x1
x2
sinx1/cosx2
child2/mutation

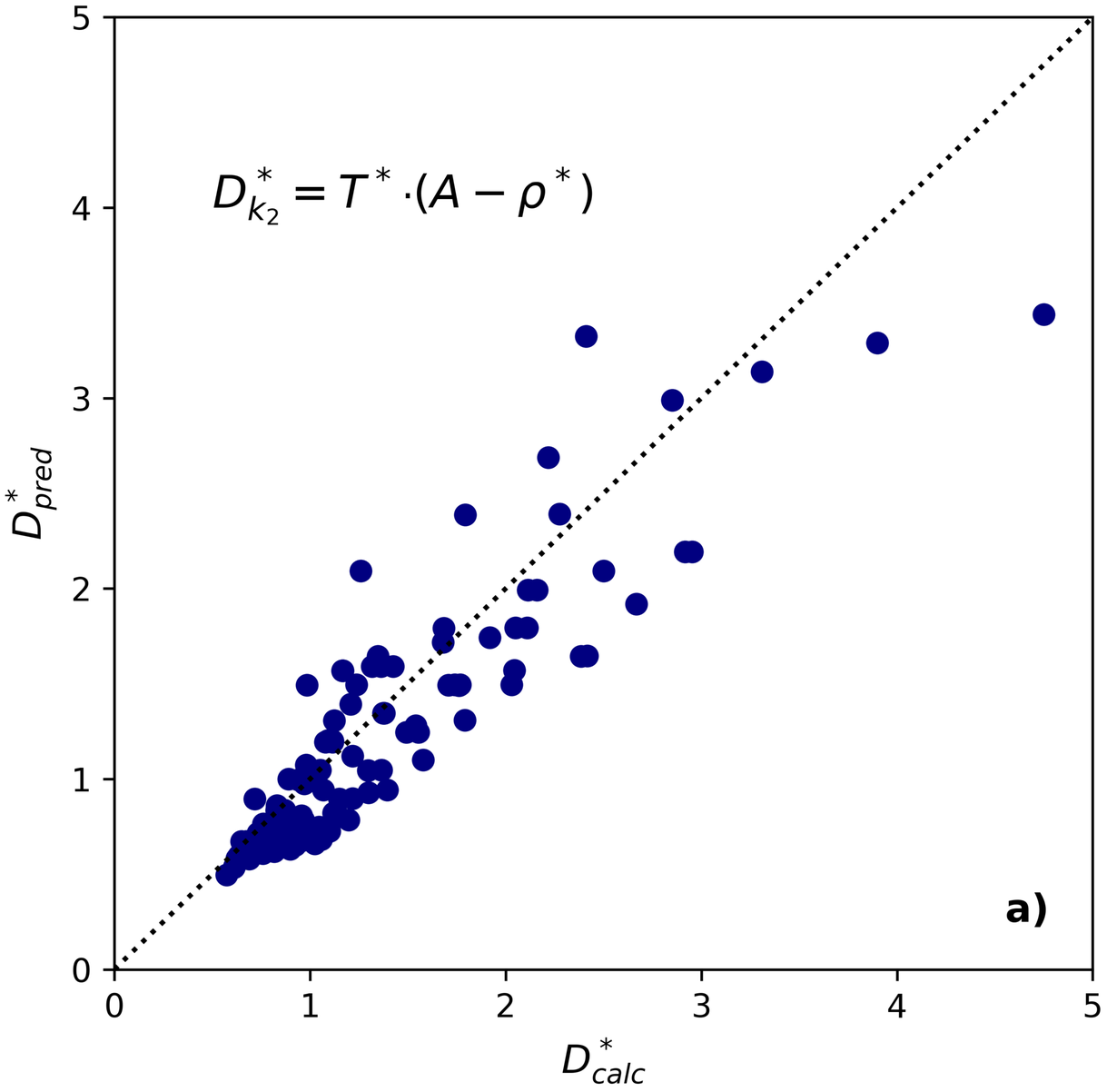

a)
G: gas
L: liquid
S: solid
SC-G
SC-L
non-rigid
SC-L
rigid
SUPERCRITICAL (SC)
Widom
line
Frenkel line
L+S
S
G+L
G+S
L
triple point
G+L+S
G
critical
point
T_c
ρ_c
T^*
ρ^*
b)
D^*
ρ^*
T^*

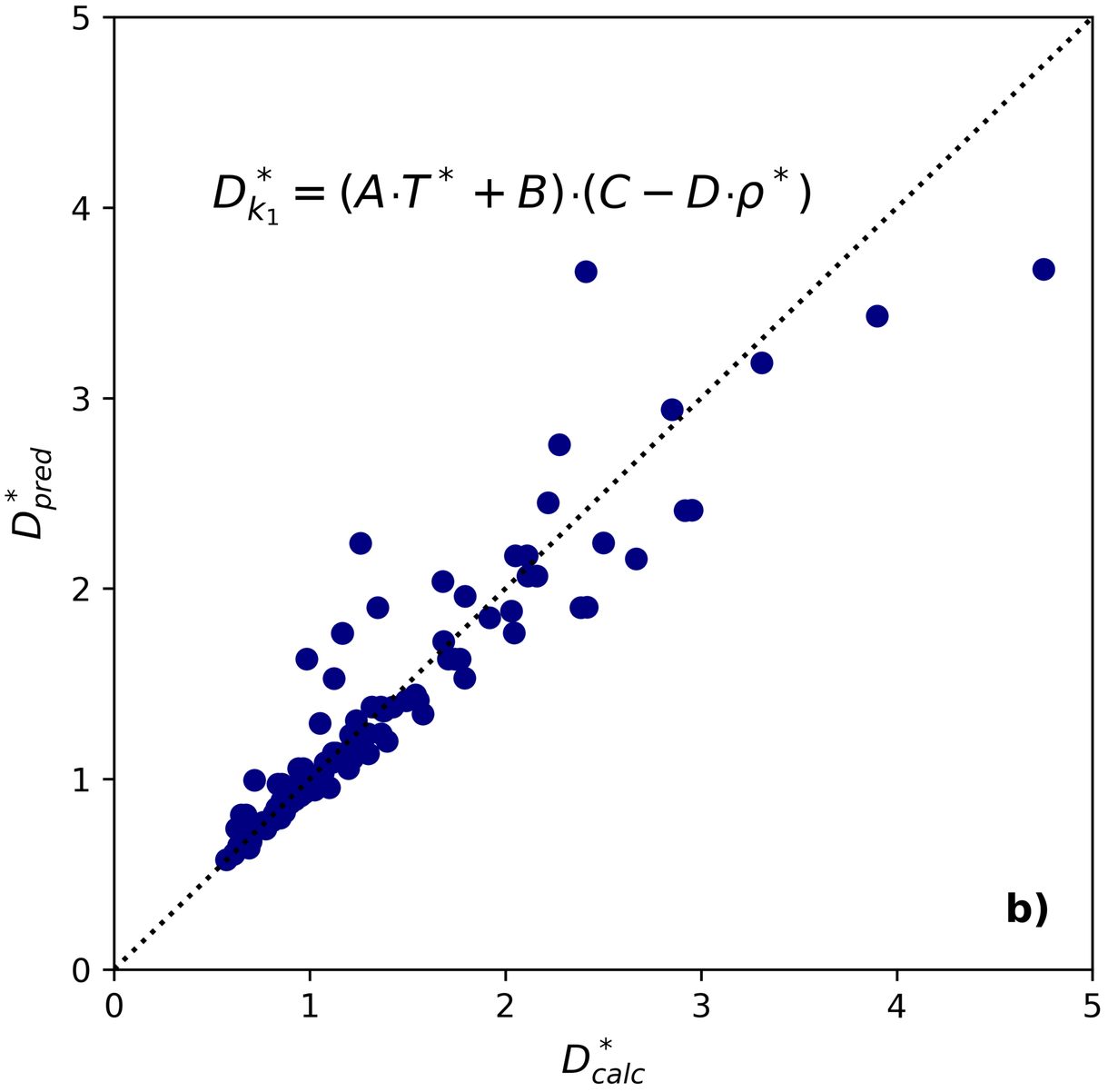

a)
b)
c)
d)
e)
f)
g)
D^*
ρ^*
T^*

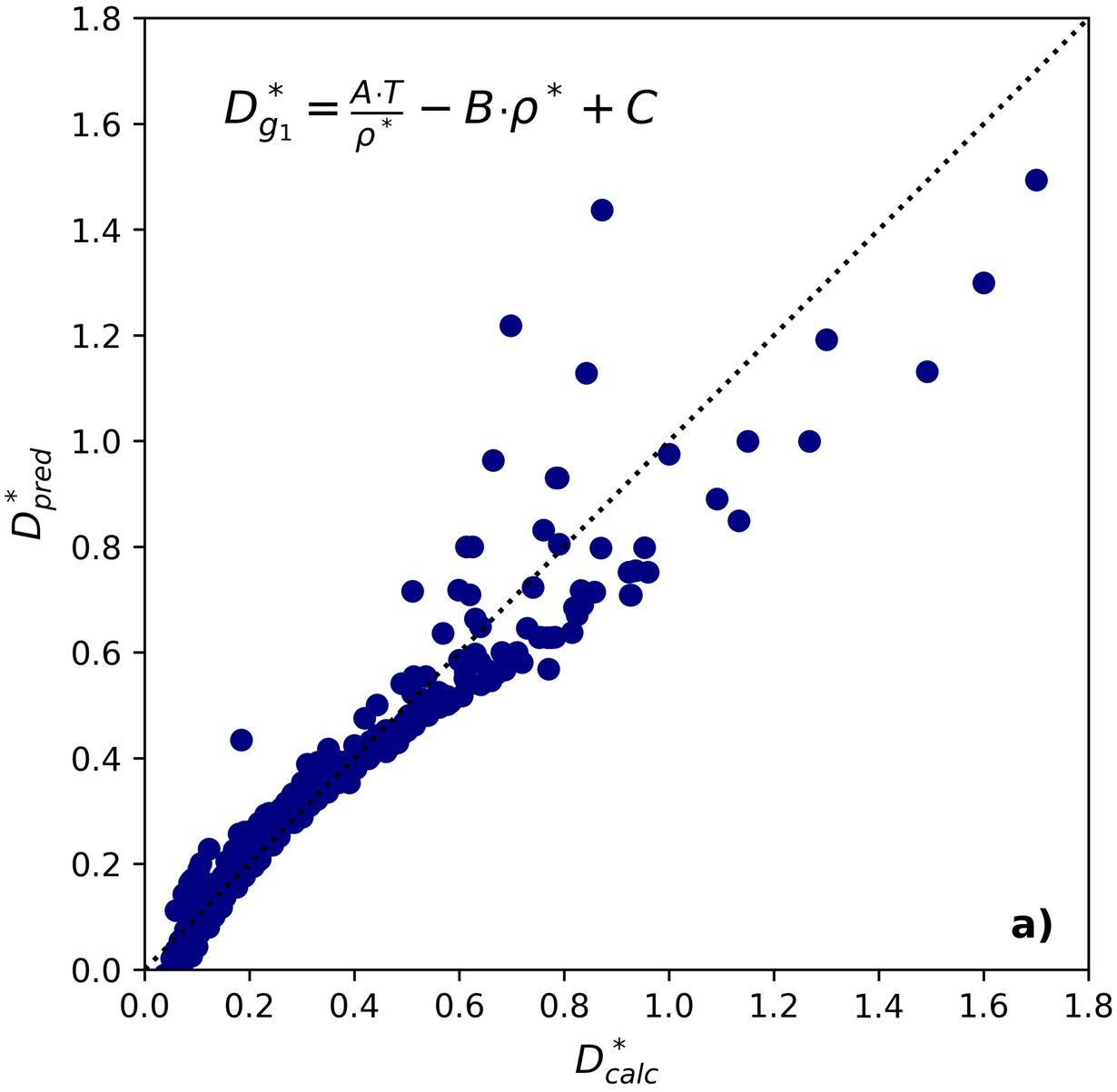

$D^*_{p_1} = \frac{A \cdot (T^* + B)}{\rho^*}$
D^*_{pred}
D^*_{calc}
0
5
10
15
20
25
30
35
a)

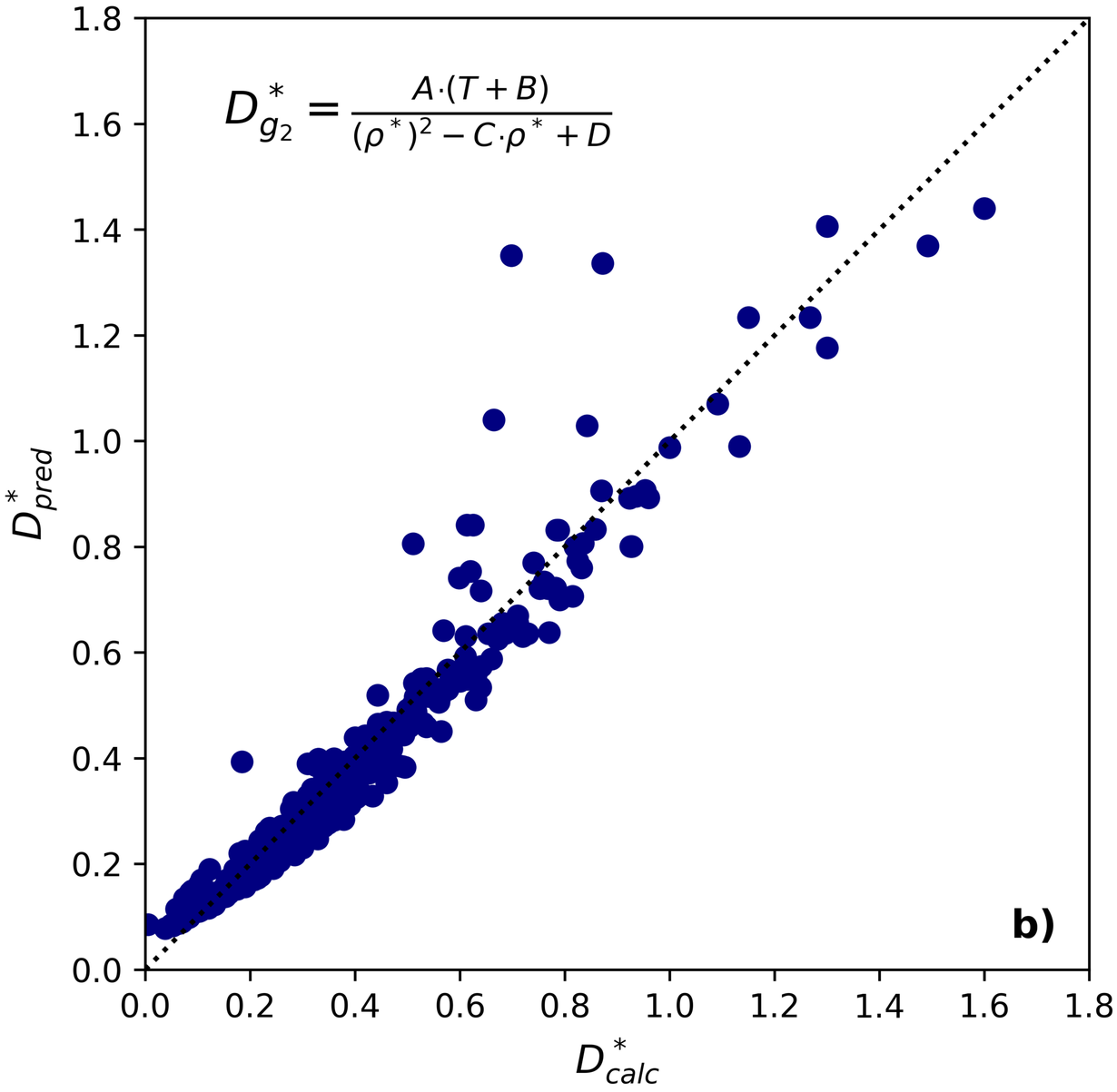

$D^*_{p_2} = \frac{A \cdot T^*}{\rho^* \cdot (T^* + B)}$
D^*_{pred}
D^*_{calc}
0
5
10
15
20
25
30
35
b)

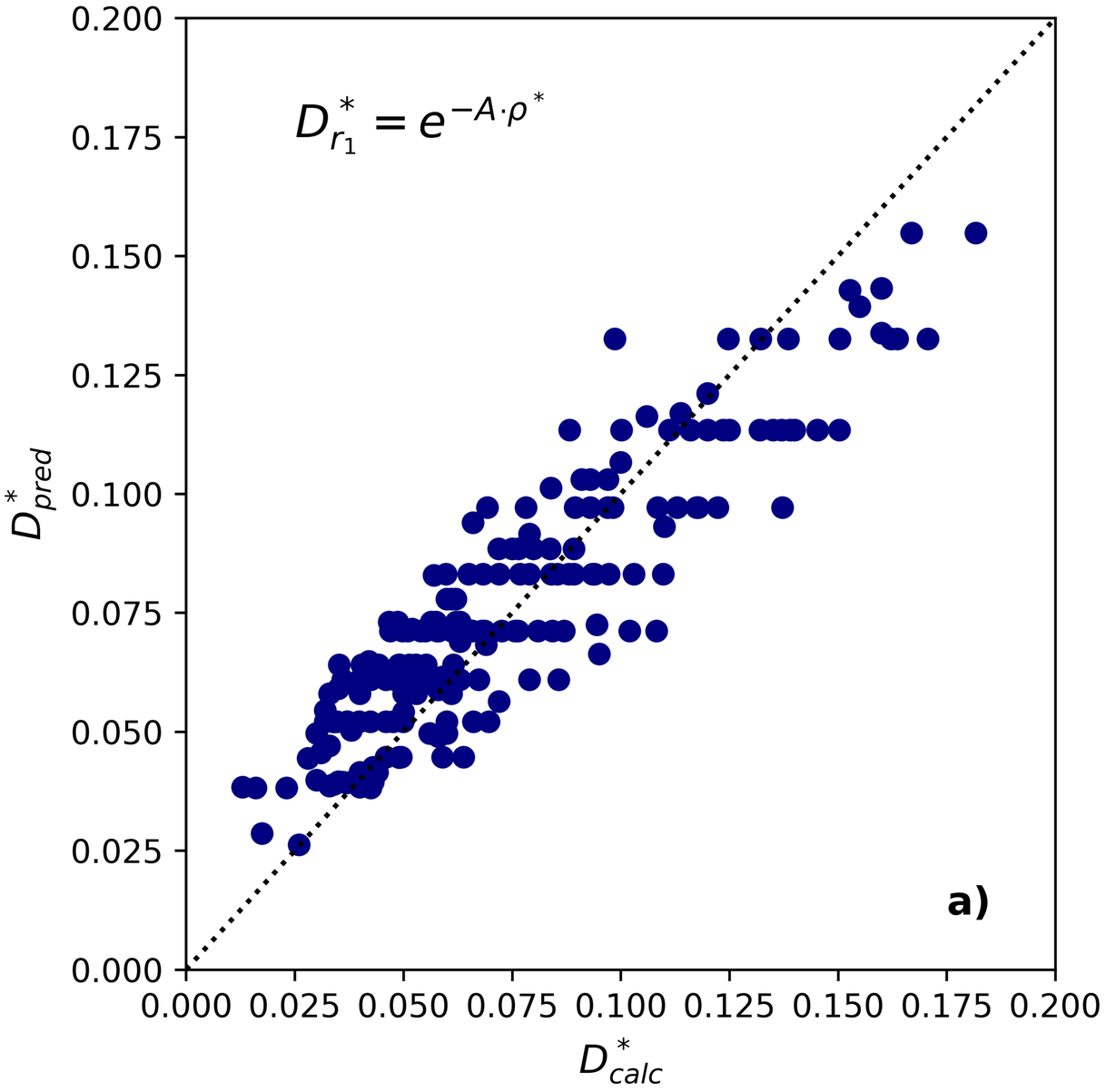

$D^*_{c_1} = \frac{T^*}{\rho^* + A}$
D^*_{pred}
D^*_{calc}
a)
$D^*_{c_2} = T^* - A \cdot \rho^* + B$
D^*_{pred}
D^*_{calc}
b)

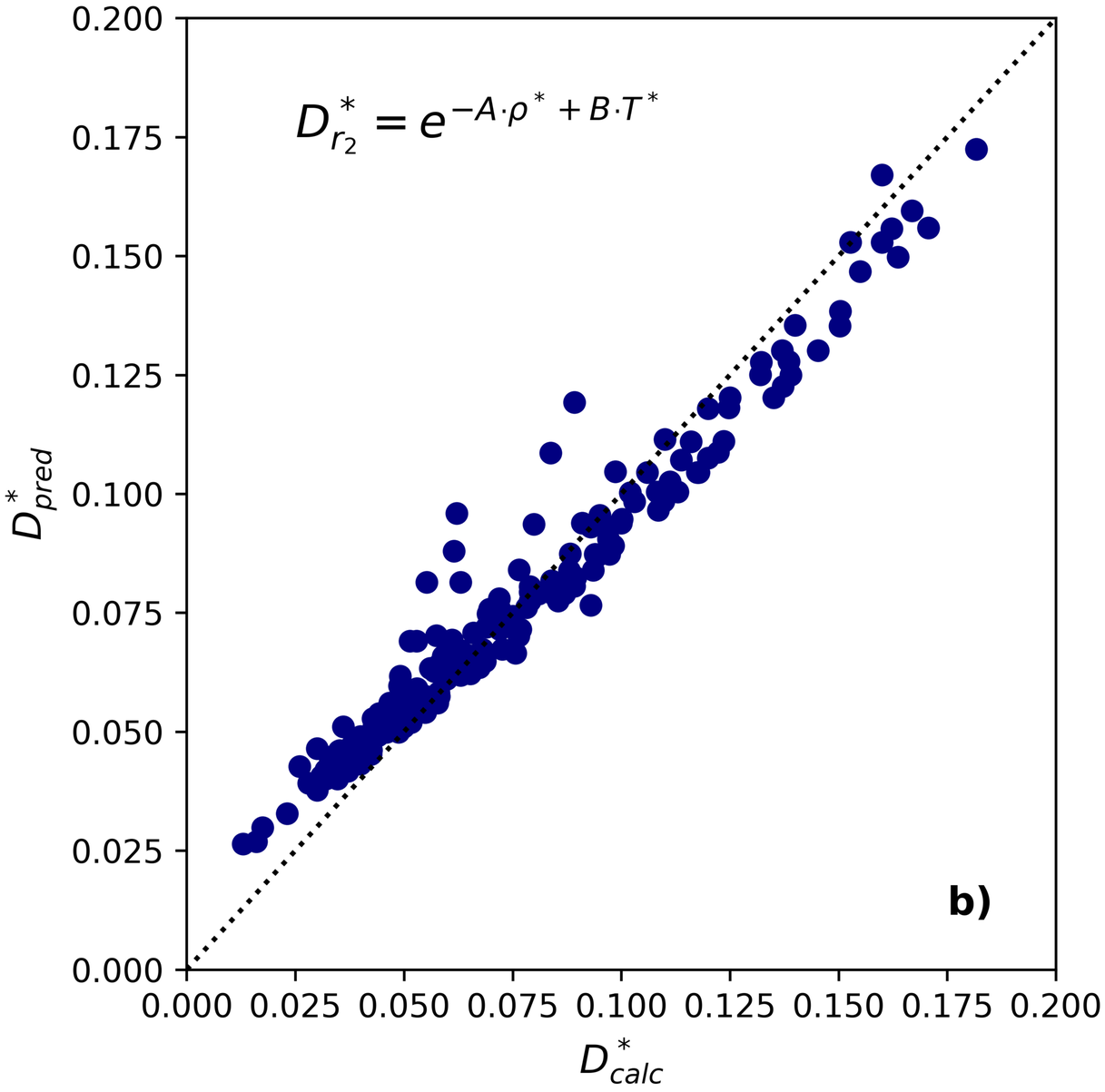

$D^*_{k_2} = T^* \cdot (A - \rho^*)$
D^*_{pred}
D^*_{calc}
0
1
2
3
4
5
a)

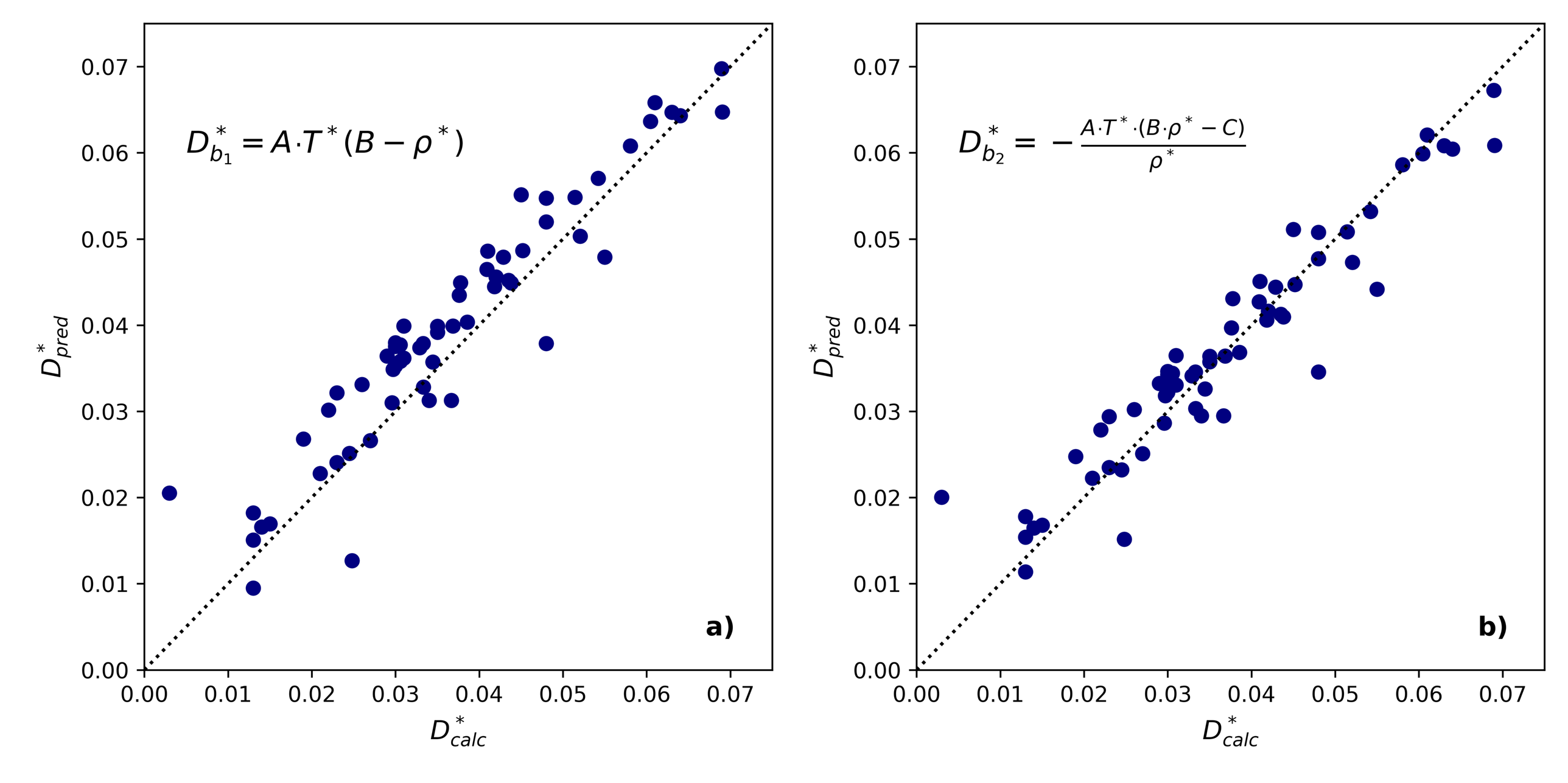

$D^*_{k_1} = (A \cdot T^* + B) \cdot (C - D \cdot \rho^*)$
D^*_{pred}
D^*_{calc}
0
1
2
3
4
5
b)

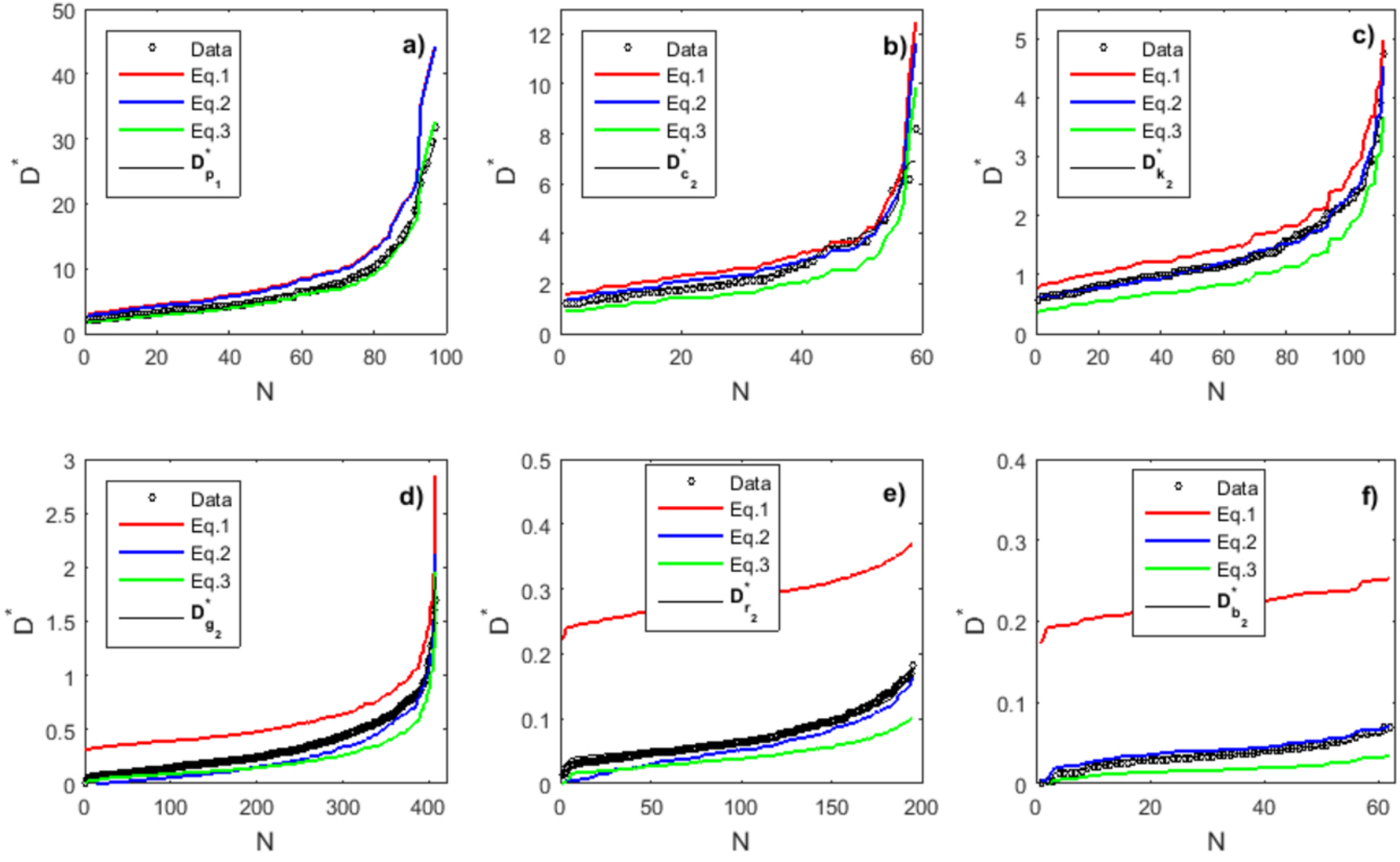

$D^*_{g_1} = \frac{A \cdot T}{\rho^*} - B \cdot \rho^* + C$
D^*_{pred}
D^*_{calc}
a)


$D^*_{g_2} = \frac{A \cdot (T + B)}{(\rho^*)^2 - C \cdot \rho^* + D}$
D^*_{pred}
D^*_{calc}
b)


$D^*_{r_1} = e^{-A \cdot \rho^*}$
D^*_{pred}
D^*_{calc}
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
a)


$D^*_{r_2} = e^{-A \cdot \rho^* + B \cdot T^*}$
D^*_{pred}
D^*_{calc}
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
b)


$D^*_{b_1} = A \cdot T^* (B - \rho^*)$
D^*_{pred}
D^*_{calc}
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
a)
$D^*_{b_2} = -\frac{A \cdot T^* \cdot (B \cdot \rho^* - C)}{\rho^*}$
b)


a)
Data
Eq.1
Eq.2
Eq.3
$D^*_{p_1}$
b)
$D^*_{c_2}$
c)
$D^*_{k_2}$
d)
$D^*_{g_2}$
e)
$D^*_{r_2}$
f)
$D^*_{b_2}$
D*
N