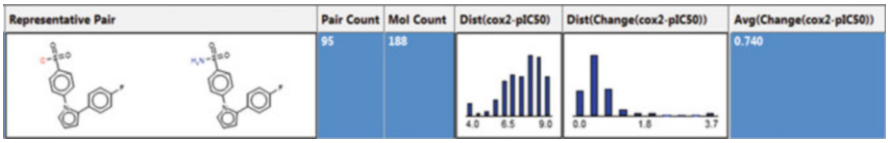

Fig. 6 Distribution plots of "cox2-pIC50" values for 188 molecules, the change in "cox2-pIC50" values for 95 pairs, and the average change in "cox2-pIC50" value for replacing CH3 with NH2

MedChem Studio then displays distribution plots for the change in value for all properties (only one property is shown here) in the spreadsheet along with other statistics such as average change in property value. An example of these distribution plots is shown in Fig. 6.

6 Building Classification and Regression Models

QSAR and quantitative structure–property relationship (QSPR) models are mathematical functions that relate molecular and atomic descriptors to biological activity or to other physical properties. Many different machine-learning algorithms can be used to create predictive models – multiple linear regression (MLR), partial least squares (PLS), artificial neural networks (ANNs), and random forest (RF), to name a few. Descriptors can be based on the atomic 3D coordinates of atoms within molecules or on the molecular connectivity alone, i.e., the 2D structure of the molecule. Our own work focuses primarily on 2D QSAR models based on artificial neural network ensembles (ANNEs) – collections of ANNs whose outputs are combined to arrive at the final result.

Regardless of the descriptors and machine-learning algorithm used, high-quality data is necessary in order to create a useful QSAR/QSPR model. One needs to thoroughly check that structures are correct and associated with the correct data value. Situations where differing experimental values exist for the same compound need to be resolved by omitting the compound, choosing a specific experimental result, or perhaps using a median or average value. It is also usually a good idea to represent the compound structure in terms of the most prevalent tautomer under the assay conditions used. Our protocol also neutralizes any ionized centers unless, like quaternary ammonium cations, they bear a permanent formal charge. Other QSAR methods may require the compounds to be in their predominant charge state at pH 7.4, e.g., a carboxylic acid would be represented as an anion. Whichever standard state is chosen, it needs to be consistently applied across the full data set as well as when generating predictions for structures not in the original data set, e.g., synthesis candidates.

The descriptors considered for inclusion in the model must be relevant to the property to be modeled. Bulk properties (logP, solubility, permeability, etc.) depend on the whole molecule. Thus, molecular descriptors can be used to accurately describe these types of properties. However, other properties such as pK_a or sites of metabolism are dependent on the atomic environment around the atom of